Biological species in the viral world

Louis-Marie Bobay[a,b,1] and Howard Ochman[a]

[a]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712; and [b]Department of Biology, University of North Carolina at Greensboro, Greensboro, NC 27402

Edited by Edward F. DeLong, University of Hawaii at Manoa, Honolulu, HI, and approved May 2, 2018 (received for review October 6, 2017)

Due to their dependence on cellular organisms for metabolism and replication, viruses are typically named and assigned to species according to their genome structure and the original host that they infect. But because viruses often infect multiple hosts and the numbers of distinct lineages within a host can be vast, their delineation into species is often dictated by arbitrary sequence thresholds, which are highly inconsistent across lineages. Here we apply an approach to determine the boundaries of viral species based on the detection of gene flow within populations, thereby defining viral species according to the biological species concept (BSC). Despite the potential for gene transfer between highly divergent genomes, viruses, like the cellular organisms they infect, assort into reproductively isolated groups and can be organized into biological species. This approach revealed that BSC-defined viral species are often congruent with the taxonomic partitioning based on shared gene contents and host tropism, and that bacteriophages can similarly be classified in biological species. These results open the possibility to use a single, universal definition of species that is applicable across cellular and acellular lifeforms.

speciation | recombination | biological species concept | gene flow | asexuality

The delineation of species, and the assignment of individuals to species, is notoriously problematic for asexual organisms (1). Like other asexual organisms, viruses reproduce clonally; but unlike cellular organisms, viruses do not possess the universally distributed genes, such as ribosomal genes and replication-related proteins, that are typically used to reconstruct molecular relationships. Also, due to their relatively small genome sizes, and their high rates and diverse modes of evolution, it has been difficult to classify viruses by a unified framework.

At the most fundamental level, viruses are grouped into broad classes according to their form of genetic material and replication mechanism (e.g., the Baltimore classification) (2); and below these classes are refinements into taxonomic ranks based on virion morphology, genome contents, host range, antigenicity, and sequence identity (3–6). Although generally conforming to a Linnaean taxonomic hierarchy (www.ictvonline.org), viral classification below the level of subfamily is fraught with inconsistencies (7). There are guidelines for classifying viruses to a species—currently defined by the International Committee on Taxonomy of Viruses (ICTV) as a "monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria" (8)—but because these defining properties are not uniformly applied, species designations can be contrived rather than having a consistent biological basis (7). Moreover, this definition of a viral species, although stipulating monophyly, remains arbitrary about the benchmark by which clades attain species status, since multiple clades at various phylogenetic depths can be extracted from a single tree.

A more pragmatic approach has been to classify viral species according to host tropism; however, this implies genetic isolation between conspecifics that infect different host species. Moreover, viral sequences extracted from metagenomic datasets often lack information about hosts or about characteristics of the virion particle, making classification based on these characteristics unfeasible. Due to these limitations, it has been advised that a genome-based classification scheme would be useful for viral taxonomy in the metagenomic era (7). Species delineation based on genomic sequence information has become standard practice for bacteria and archaea (9), and similar taxonomic schemes can be applied to viruses, although different criteria might be needed for dsDNA, ssDNA, and RNA viruses to account for their diverse genome structures and rates of evolution.

To circumvent the demands of phylogenetic analysis and the frequent lack of species-defining characteristics, many species of viruses have been delineated using sequence identity thresholds, an approach that is commonly applied to asexual lifeforms but is often criticized on account of the arbitrary nature of such thresholds (10–12). Because viruses, as a whole, span a wide range of mutation rates, it is not possible to specify a particular sequence identity threshold that might correspond to species, as has been suggested for bacteria (13). Current classifications of viral species based on sequence similarity are varied and rely on clade-specific thresholds (11), preventing the emergence of a unified species definition in viruses. However, if viruses engage in homologous gene exchange, either by encoding their own recombination system or by recruiting their host's (14, 15), it is possible that they form true biological species analogous to those defined in sexual cellular organisms.

Despite their asexual mode of reproduction, several microorganisms engage in sufficient levels of gene flow (i.e., homologous recombination) to distinguish biological species (16–21). We recently introduced a methodology that uses genome sequences to delineate reproductively isolated groups within bacteria and archaea based on their barriers of gene exchange, thereby offering a single framework to define biological species for any set of recombining organisms (21). Similarly, viruses do not rely on sexual reproduction; however, recent analyses of homologous recombination and network reconstructions suggest that species delineation based on gene flow would be possible for bacteriophages infecting cyanobacteria (22–26). Moreover, an experimental evolution study has shown that the mechanisms of

Significance

The biological species concept (BSC) has served as the basis for defining species for over 75 years. Members of a biological species are defined by their ability to exchange genetic material, and it was originally thought that asexual lineages were not amenable to species-level classification based on the BSC since clonal individuals are reproductively isolated from one another. In this study, we demonstrate that the rates and patterns of gene exchange in acellular organisms (viruses and bacteriophages) allow the assignment of true biological species, an essential step to organizing the tree of life. Our results show that a universal species definition, based on the BSC, can be used to define biological species in all major lifeforms.

Author contributions: L.-M.B. and H.O. designed research; L.-M.B. performed research; L.-M.B. analyzed data; and L.-M.B. and H.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: ljbobay@uncg.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717593115/-/DCSupplemental.

Published online May 21, 2018.

speciation occurring in sexual organisms might also operate in viruses (27).

Although the routes and rates of homologous recombination remain poorly described for most viruses, we show that viruses with similar gene contents—typically those classified as members of the same genus—frequently recombine and that gene exchange occurs readily among highly divergent orthologs. Because so few viral taxa were found to evolve clonally, viruses and bacteriophages, like cellular organisms, can be classified into true biological species in accordance with the Biological Species Concept, and that their species limits are largely defined by their host tropism and overall genome contents.

Results

Defining True Biological Species. Because members of biological species are characterized by their capacity for gene exchange, we assessed the degree of recombination within eight genera of animal viruses for which there were both large numbers of fully sequenced genomes ($n \geq 15$) and core genomes of sufficient size (≥20 kb) to accurately determine whether polymorphic sites arose by mutation or recombination. Our analyses identify gene flow between homologous sequences (i.e., recombination, homologous gene exchange) but does not infer events of horizontal gene transfer (i.e., gene acquisition), which need not involve homologous exchange. In some instances, genomes from a particular genus were excluded from the analysis due to their low amounts of shared gene content with other members of the genus (*Methods* and *SI Appendix*, Fig. S1). To estimate the prevalence of gene flow (i.e., homologous recombination) within these redefined genera, we estimated the ratio of homoplasic (*h*) to nonhomoplasic (*m*) polymorphisms along the core genome of each genus. Homoplasies are polymorphisms that are not compatible with vertical inheritance from a single ancestral mutation and likely result from the exchange of alleles through homologous recombination. High *h*/*m* ratios are indicative of a substantial signal of gene flow, and low *h*/*m* ratios are indicative of clonal or nearly clonal evolution (Fig. 1, *Top*). Sharp variations in the *h*/*m* ratios computed across different combinations of genomes—as detected by the exclusion criterion (*Methods*)—can indicate the presence of one or more genomes that do not engage in gene flow with the rest of the population. In all cases, viruses classified to the same genus showed a signal of gene flow (Fig. 1), and there was no statistical evidence that some viruses are "sexually" isolated from other members of the genus. These results suggest that many ICTV-defined viral genera are actually true biological species, as defined by the BSC.

We extended this analysis to bacteriophages infecting *Mycobacterium smegmatis*, which represent the largest sampling of bacteriophage genomes infecting a single host and which have been classified into multiple clusters and subclusters based on gene content (28, 29). By applying the same procedures, we observed that all but one of the 17 bacteriophage clusters (cluster C1) are compatible with a BSC-like definition of species (*SI Appendix*, Fig. S2). Within cluster C1, we determined that one bacteriophage (*Tonenili*) is not recombining with the other members of the cluster, but after excluding this genome from our analysis, we retrieved a signal of homogeneous gene flow within the cluster (*SI Appendix*, Fig. S3). It should be noted that the bacteriophages in our dataset have diverse origins and analyzing viruses from similar geographic and ecological locations would likely result in tighter networks of gene flow, as suggested by previous studies (22–26).

These results suggest that gene flow significantly impacts viral microevolution, as has been reported for cyanophages (22–26). To further visualize the impact of gene flow on viral and bacteriophage evolution, we built dendrograms with SplitsTree, which depicts events of homologous recombination on a phylogenetic network (*SI Appendix*, Figs. S4 and S5). As predicted by the *h*/*m* ratios estimated from the core genomes of these taxa, most phylogenetic networks show high levels of reticulate evolution, indicative of a gene flow among members of a taxon.

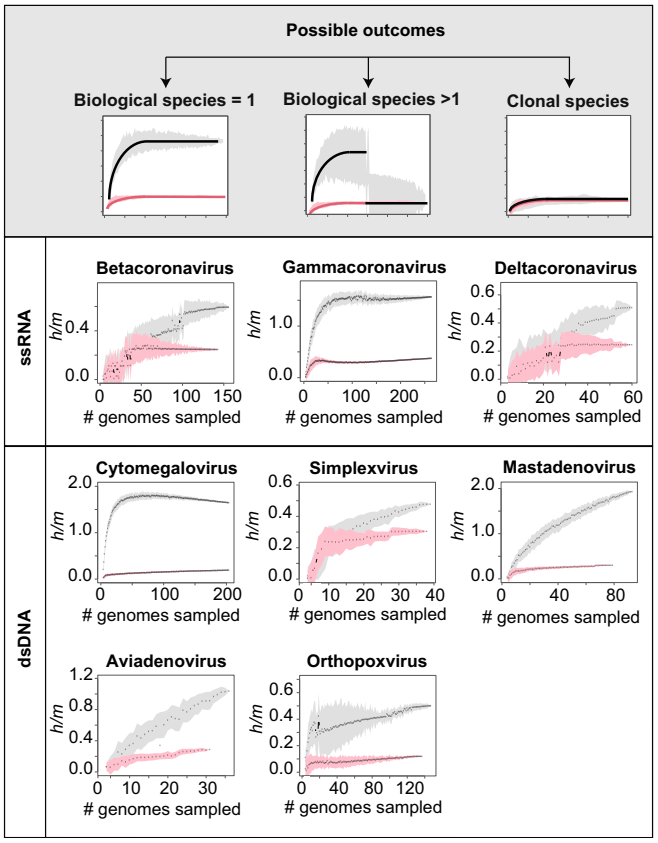

Fig. 1. Recognizing biological species. In each set of genomes comprising *n* strains, nonredundant combinations of *i* strains (ranging from 4 to $n - 2$ strains) were subsampled 100 times for each value of *i*. At each iteration of subsampling, the *h*/*m* ratio—the ratio of polymorphisms attributable to homoplasy relative to those attributable to mutation—was calculated for the concatenated alignment of genes common to all strains. Within the bivariate plots, black dots are medians and the gray-shaded region is the SD of the indicated number of subsampled combinations of strains. Red dots and pink-shaded regions denote median *h*/*m* values and SD 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. *Top* shows the three possible outcomes of these analyses: when there are no barriers to gene flow among strains (*Left*); when a discontinuity is produced by inclusion of a strain that does not participate in gene exchange (*Center*); and when there is clonal evolution or an absence of gene flow (*Right*). *Lower* displays results obtained for dsDNA and ssRNA viruses.

Effects of Sampling and Evolutionary Rates on Defining Viral Species. A key challenge for defining viral species based on gene flow is that viruses typically evolve at very fast rates (especially RNA and ssDNA viruses) (30), which could affect the inference of recombinant sites. It is possible for homoplasies to be introduced by independent, convergent mutations, which would result in an overestimation in the number of polymorphic sites that are introduced by recombination. To test for the effects of convergent mutations on the generation of homoplasies, we simulated sequences that mimicked the features of each viral dataset with respect to level of polymorphisms, nucleotide composition, and tree topology.

The large majority of viral clades and bacteriophage clusters that we defined as species displayed higher *h*/*m* ratios than the simulated sequences, indicating that convergent mutations represent only a small fraction of the homoplasies. However, three viral genera (*Simplexvirus*, *Betacoronavirus*, and *Deltacoronavirus*) and three bacteriophage clusters (A6, A9, and N) did not diverge

substantially from random expectation. Such patterns can be caused by the lack of recombination but can arise also from the inclusion of multiple sexually isolated subclades, which would reduce the overall signal of recombination (21).

To test whether the three viral genera recognized as clonal (*Simplexviruses*, *Betacoronaviruses*, and *Deltacoronaviruses*) were actually composed of multiple recombining subclades, we partitioned the members of each genus into subclades based on their phylogenetic relationships. These subclades were built by progressively eliminating external taxa, and we then evaluated the pattern of recombination among the members within each of the newly assembled subclades. After decomposing these genera into smaller ensembles, we recovered clear signals of gene flow for *Simplexvirus* and *Deltacoronavirus* (*SI Appendix*, Fig. S6), indicating that these two genera do not evolve clonally, but rather, each contains multiple biological species that do not recombine with one another. Only a single viral genus, the *Betacoronaviruses*, appeared evolving in a clonal—or nearly clonal—manner. These results show that having similar gene contents is not the only condition for viruses to engage in gene flow and indicate that other forces—such as ecological factors—might restrict the ability of viruses to recombine.

Since recombination between viruses can only take place within coinfected cells, we reasoned that the degree of host tropism might affect the delineation of viral species, such that biological species might comprise viruses that infect a single host species. Despite infecting a wide range of mammalian and bird species, *Mastadenoviruses*, *Aviadenoviruses*, and *Orthopoxviruses*, were each determined to be a single biological species, suggesting that they switch hosts frequently enough to allow opportunities for recombination or that any interruption of gene flow is too recent to be detected.

In contrast, two viral genera that were originally viewed as clonal but later redefined as containing multiple biological species (*Simplexviruses* and *Deltacoronaviruses*) initially displayed a wide host tropism: the analyzed strains of *Simplexviruses* were isolated from humans, monkeys, bats, and rabbits; and *Deltacoronaviruses* from pigs and several species of birds. After redefining biological species in both of these genera, we observed that each of the newly defined species is confined to a single host species. The redefined biological species of *Simplexviruses* is confined to strains infecting human hosts, and the redefined biological species of *Deltacoronaviruses* is confined to strains infecting pigs. In contrast, the *Betacoronaviruses*, the only genus for which we were unable to define a single biological species, comprise viruses isolated from humans and camels, indicating that this genus might include two biological species of different host tropisms; but its sample size was too small to test this hypothesis.

Recombining Viruses Are Highly Promiscuous. We analyzed the degree of genomic divergence among members in each of the species that we delineated and observed that despite potential sampling biases, several species contain highly divergent members (Fig. 2). For example, *Mastadenovirus* contains members that share only 58% sequence identity for genes constituting their core, despite the clear signal of gene flow among members. Such cases suggest that viruses are able—directly or indirectly—to engage in homologous recombination despite high levels of sequence divergence.

To compare the boundaries of viral species defined by the BSC to those designated by an alternative genome-based method, we estimated the average nucleotide identity (ANI) along the core genome for members of the same biological species and then calculated the number of groupings (ANI species) at different sequence-identity thresholds. (This analysis did not assume a single reference genome and it grouped together any pair of genomes that had a higher ANI value than the given threshold.) At sequence-identity thresholds that are commonly employed, there are often large numbers of ANI species within a single BSC-defined species (*SI Appendix*, Fig. S7), owing to the fact that

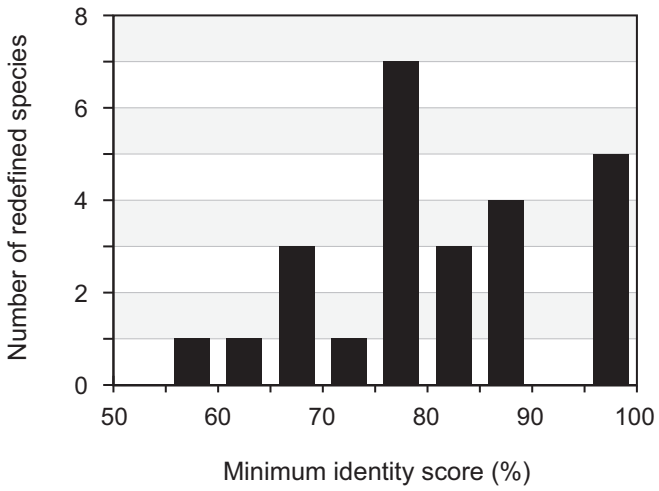

Fig. 2. Maximum sequence divergence between members of viral and bacteriophage species. Shown are average nucleotide sequence identity values for orthologous genes shared by the two maximally divergent strains within each biological species.

divergent genomes retain the ability for homologous exchange. Naturally, the number of ANI species depends on the selected threshold; but importantly, each viral and bacteriophage taxon is impacted differently when analyzed at alternate sequence thresholds. For example, bacteriophage cluster E forms a single ANI species regardless of the chosen threshold, whereas cluster A2 constitutes one, or up to 30, ANI species depending on the threshold (*SI Appendix*, Fig. S7). Similarly, *Cytomegalovirus* would constitute a single ANI species, whereas *Mastadenovirus* would comprise ≥20 species if applying the same ANI threshold (*SI Appendix*, Fig. S7). Therefore, sequence thresholds do not define cohesive populations and—in many cases—distinct ANI species engage in gene flow.

The ability of divergent viruses to recombine can be associated with the presence of their own recombination systems, which are often more permissive than those of the host (31). A previous study identified recombinase genes across a large set of bacteriophages (14), and three of these genes, Bxz1 *gp201*, 244 *gp117*, and PMC *gp61*, were identified in bacteriophage species C1, E, and F1, respectively. These are, in fact, the three species that displayed the highest rates of recombination among the bacteriophage clusters we tested (*SI Appendix*, Fig. S2), supporting the view that phage-encoded recombinase genes promote high rates of genomic diversification through permissive homologous recombination.

In light of the potential for recombinational exchange between highly divergent sequences, we also explored whether members of different bacteriophage clusters exchange genes and actually constitute a single biological species. By examining those bacteriophage clusters that share high numbers of homologous genes (e.g., clusters A2, A6, and A9, and clusters A3 and A4), we observed that certain clusters engage in gene flow with one another (*SI Appendix*, Fig. S8). However, the pattern of gene flow among some of these bacteriophage clusters is asymmetric: members of cluster A2 are able to capture DNA from members of A6 and A9, but members of clusters A6 and A9 do not recombine with one another. This situation is analogous to so-called "ring species" in animals (32): clusters A2, A6, and A9 could be viewed as a single biological species, although some subpopulations (i.e., A6 and A9) are sexually isolated from one another. It is also possible that this asymmetry in gene flow is affected by sampling biases, since these bacteriophages were isolated from different locations.

Discussion

By analyzing patterns of recombination among members of named viral genera and clusters of bacteriophages, we show that the Biological Species Concept, i.e., the delineation of species based on the ability for gene flow among members, can be extended to include viruses and bacteriophages. Despite the limited number of viral clades for which there is sufficient genomic information for analysis, our results show that taxonomic boundaries based on gene flow can be established in viruses, as previously reported for bacteriophages infecting cyanobacteria (22–26) and in bacteriophage lambda that have coevolved with their host (27), contrasting traditional views that viruses, being asexual, evolve clonally. Although homologous recombination has been reported for many viruses, including the genera analyzed in this study (33–39), its frequency and significance on species delineation has been underestimated. In contrast to bacteria, viruses and bacteriophages have high rates of mutation, genetic reassortment and exchange, and gene uptake (especially for RNA and dsDNA viruses) (30), which can generate substantial genetic and genomic diversity. However, viral recombination is not indiscriminate but is confined largely to entities with similar gene contents, such that several of previous groupings based on shared gene contents actually constitute biological species.

Viruses can be organized into a hierarchical classification scheme based on genome contents and overall sequence similarity (6, 26, 40–45); however, genomic features are not the only attributes that dictate species membership. Because homologous exchange requires access to conspecifics, the boundaries of certain viral and bacteriophage species are also imposed by host range or tropism. Therefore, it is necessary to integrate species-level classifications based on gene flow with network-based approaches for higher taxonomic ranks to produce an accurate and consistent classification scheme for viruses and bacteriophages.

Our delineation of viral species is based on recognizing gene flow between viruses, here defined as exchange between gene homologs. However, viruses and bacteriophages engage in additional and more complex exchange processes in which genes are gained or lost through events of horizontal transfer. Acquisition of new genetic material can lead to genome mosaicism, with the result that very divergent viruses might share modules of highly similar contents and sequence (46). Notoriously, such genome mosaicism led to taxonomic conflict in lambdoid bacteriophages, since some possess gene modules of nearly identical sequence (and potentially engage in gene flow) while others lack these sequences (47).

Genome mosaicism has mainly hampered the taxonomy of temperate bacteriophages (48), which are thought to be prone to events of horizontal transfer due to their stable existence in the host chromosome and exposure to other infecting bacteriophages. Although high levels of gene acquisition will disrupt hierarchical schemes of classification (49), horizontal transfer seems to affect relatively few genes in cellular organisms and in most viruses (26, 48), such that a set of core genes can be evaluated for homologous exchange. And in fact, gene-sharing networks suggest that temperate bacteriophages can be ordered into a hierarchal taxonomic framework despite their mosaic structure (26).

That viruses and bacteriophages can be classified into species based on a biological process—and on the same biological process used for cellular organisms—allows the mechanics of cladogenesis to be compared across all lifeforms. First, we can evaluate the adequacy and applicability of the BSC for species-level classification within each of the major groups of organisms. One shortcoming of the BSC has been its inability to delineate species in asexual organisms, which are ubiquitous among prokaryotes but usually considered to be rare in animals, plants, and fungi (50). Despite differences in cellularity and reproductive processes, the overwhelming majority of lineages in every group can be classified into species based on the BSC (Fig. 3). The BSC was originally formulated for animals, in which parthenogenesis and self-compatibility are rare, and its application to plants was somewhat more limited because it was originally estimated that upwards of 20% of plants reproduced solely by selfing (51, 52). However, reevaluation of the botanical literature suggests that outcrossing is as common in seed plants as in animals and that only about 10% of ferns do not regularly outcross (50). The ability to reproduce clonally is more common in protists (50), but evidence suggests that the extent of sexual reproduction—or some related mechanisms—remains largely underestimated in these lineages (53–55). Although bacteria are asexual sensu stricto, only about 15% are not amenable to classification by the BSC (21), and similarly, less than 20% of bacteriophage and viral groups do not engage in sufficient genetic exchange to be classified as biological species.

The boundaries of viral and bacteriophage species are much broader than those in other groups. Whereas the maximal divergence between members of animal and plant species is rarely generally close to 1%, and rarely greater than 5% [as tabulated for neutral sites (56–58)], in every viral species that we examined, there are conspecifics whose homologs differ in sequence by over 20% (Fig. 2). In general, viral species accommodate much higher levels of sequence divergence than bacteriophage species, and all but three of the bacteriophage species (B2, B3, and E) show less than 5% sequence difference between their maximally divergent members (but note that all analyzed bacteriophages infect the same host: *M. smegmatis*). This ability to recombine with very distant homologs even exceeds what has been reported for bacteria, in which 30% of BSC-defined species have a maximal divergence among conspecifics greater than 5% (21).

The ability for highly divergent viruses to engage in gene flow likely results from their high rates of mutation coupled with action of diversifying selection, since hosts typically impose strong selective pressures that promote diversification (59, 60). Compared with cellular organisms, viruses are capable of recombining more highly mismatched sequences (31); and consistent with this hypothesis, we note that bacteriophages encoding their own recombination systems have the highest rates of gene flow. That viruses can be classified into species based on gene exchange is somewhat paradoxical, since recombination is only rarely required to accomplish viral lifecycles (15, 61), and the majority does not encode for recombination enzymes or rely on this mechanism to propagate. In such cases, recombination can be achieved through using the host recombinase or through "copy choice" processes, in which hybrid genomes are generated by template disassociation and reassociation of the replication complex in coinfected cells (62–64).

Finding that viruses and bacteriophages are amenable to species-level classification based on the Biological Species Concept implies that they can be fully organized in the Linnaean hierarchy established for cellular organisms. We find that biological species of viruses and bacteriophages can be associated with host tropism; however, the high incidence of host switching and uncertainties about primary host reservoirs, combined with the potential for recombination between very divergent genomes, contravenes the attachment of viral species' names to a particular host (65, 66). Moreover, linking viruses to hosts has become particularly problematic in metagenomic studies, since sequences often originate from sources unrelated to host infection and many of the traits typically used for viral classification remain inaccessible (67, 68). Although our analyses are currently restricted to fully sequenced genomes, there has been substantial progress in the recovery of identification and assembly of viral sequences from metagenomic datasets (69), thereby extending the applicability of our approach. By defining the fundamental units of viral taxonomy, our approach constitutes a step toward the consensus reached by the ICTV (7), stating that a sequence-based approach, centered on a uniform and universally relevant criterion, represents the optimal tool for classifying viruses and bacteriophages into a coherent structure.

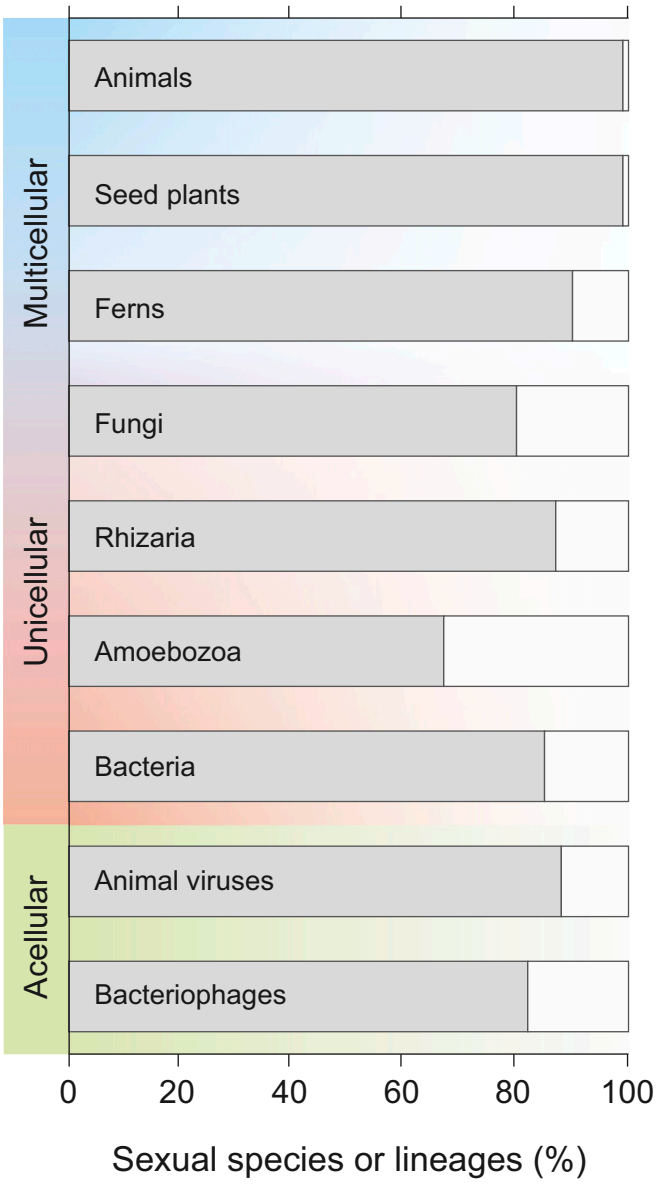

Fig. 3. Prevalence of sex and related mechanisms in cellular and acellular taxa. Frequency of sexual species was compiled by ref. 50 for animals, seed plants, ferns, and fungi and by ref. 21 for bacteria ($n = 91$) and in the present study for animal viruses ($n = 8$) and bacteriophages ($n = 17$). Values reported for Rhizaria ($n = 15$) and Amoebozoa ($n = 15$) refer to the frequency of lineages containing sexual species reported in ref. 55.

Methods

Viral Datasets and Defining Core Genomes. We retrieved 3,107 genomes of animal viruses representing 15 named genera from the RefSeq database (https://www.ncbi.nlm.nih.gov/) and the entire collection of 1,154 bacteriophage genomes infecting *M. smegmatis* (phagesdb.org) in May 2017. Animal viruses were selected based on two criteria: (*i*) the availability of a large number ($n \geq 15$) of genomes classified under the same genus name and (*ii*) an average genome size exceeding 20 kb for members of a genus (*SI Appendix*, Fig. S1, step 1). Bacteriophages were annotated in GLIMMER v3.02 (70) and grouped taxonomically into clusters based on gene content, as defined in refs. 28 and 29. For each viral genus and bacteriophage cluster, we obtained the set of orthologous proteins shared by each pair of genomes with USEACH Global v5.2 with 70% identity and 80% length conservation (71) and then defined a "core" genome as the set of single-copy orthologous genes shared by at least 85% of its members.

In cases where there were large differences in genome contents among members of a genus, we systematically redefined the genus borders as follows: For each pair of genomes, we defined the ratio of shared orthologs (S) between genomes A and B as $S_{AB} = O_{AB}/\min(A,B)$, where O_{AB} represents the number of orthologs shared by genomes A and B, and $\min(A,B)$ represent the total number of genes of the smaller genome (72). From the matrix of scores S, we grouped genomes in MCL v14-137 (73) with an inflation parameter of 1.2 (i.e., the minimal value for obtaining large clusters) and redefined each genus as the genomes included in the largest cluster. Following this procedure, a total of four genera were redefined, and their corresponding core genomes were based on each of the redefined sets of genomes (*SI Appendix*, Fig. S1, step 2). All cases in which the size of the core genome was less than 20% of the average genome size were excluded, leaving a total of the following eight viral genera: *Cytomegalovirus*, *Simplexvirus*, *Mastadenovirus*, *Aviadenovirus*, *Orthopoxvirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* (*SI Appendix*, Fig. S1, step 3). None of the bacteriophage clusters (A1–A6, A9, B1–B3, C1, E, F1, J, K1, L2, N) was subdivided, since each was characterized by a sufficiently large core genome. The lists of analyzed viruses and bacteriophages are available in Datasets S1 and S2, respectively.

For each viral genus or bacteriophage cluster, the protein sequences of each gene in the core genome were aligned with MAFFT v7.271 (74), reverse translated into their corresponding nucleotide sequences, and merged into a single concatenate. Additionally, we attempted to build additional core genomes for pairs of viral genera and of bacteriophage clusters that had similar genome contents, with the result that four pairs of bacteriophage clusters (A2–A6, A2–A9, A6–A9, and A3–A4) possessed relatively well-conserved joint core genomes consisting of 16 or more shared orthologs.

Analysis of Gene Flow. For each viral genus and bacteriophage cluster, estimates of recombination were based on homoplasies as in ref. 21. First, a matrix of core genome distances D was built for each genus or cluster using RAxML v8.2.7 (75) under a generalized time-reversible (GTR) model, and redundant genomes—those with no or little divergence ($D \leq 0.00005$) to another genome—were randomly removed (*SI Appendix*, Fig. S1, step 4). For each core genome, polymorphic sites were inferred as homoplasic when $max(D_{11}) > min(D_{10})$, where $max(D_{11})$ represents the distance between the genomes harboring the minor allele, and $min(D_{01})$ represents the minimal distance between genomes harboring a minor allele and a major allele. We then computed the ratio h/m, defined as the ratio of homoplasic (h) alleles to nonhomoplasic (m) alleles for multiple combinations of genomes, such that groups of genomes with higher h/m ratios have more polymorphisms attributable to recombination. For each viral genus or bacteriophage cluster, we randomly sampled 100 nonredundant combinations of genomes for different numbers of genomes (from 4 to $n - 2$, with n the total number of genomes in the genus or cluster being analyzed). Within each viral genus or bacteriophage cluster, we identified genomes that led to a sharp reduction of the h/m ratio relative to other genomes by applying an exclusion criterion as in ref. 21. Such genomes, as they do not recombine with other members of the population, are not considered members of the same biological species (*SI Appendix*, Fig. S1, step 5). We also applied our exclusion criterion to those pairs of bacteriophage clusters (A2–A6, A2–A9, A6–A9, and A3–A4) with highly similar core genomes. In these cases, those pairs of clusters that did not display a significant decrease in gene flow based on our exclusion criterion were considered to be the same biological species (*SI Appendix*, Fig. S1, step 6).

Simulations. We assessed the expected number of homoplasies that might be introduced by convergent mutations in each of the datasets through simulations as in ref. 21. We built a maximum likelihood tree for the core genomes in each genus or cluster with RAxML v8.2.7 (75) under a GTR model. Then, using SeqGen v1.3.3 (76), the resulting tree was applied to generate an alignment that maintained the nucleotide composition, the number of genomes, and the length of the alignment. Because phylogenetic inference considers recombination events as multiple independent mutation events (thereby overestimating the number of mutations that accumulated in the simulated alignments), we rescaled the length of the branches of the trees uniformly to match the level of polymorphisms in the simulated alignments and the real data. Each simulated alignment was then subjected to the same resampling strategy to detect homoplasies.

Phylogenetic Networks. For each viral genus and bacteriophage cluster, phylogenetic networks were built with SplitsTree v4 (77) with default parameters.

ACKNOWLEDGMENTS. We thank Kim Hammond for help with preparation of the figures. This work was supported by the National Institutes of Health award R35GM118038 (to H.O.).

1. Shapiro BJ, Polz MF (2015) Microbial speciation. *Cold Spring Harb Perspect Biol* 7: a018143.
2. Baltimore D (1971) Expression of animal virus genomes. *Bacteriol Rev* 35:235–241.
3. Abedon ST, Calendar RL (2005) *The Bacteriophages* (Oxford Univ Press, Oxford), 2nd Ed, p 768.
4. Krupovič M, Bamford DH (2010) Order to the viral universe. *J Virol* 84:12476–12479.
5. Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (2005) *Virus Taxonomy: VIIIth Report of the International Committee on Tanoxomy of Viruses* (Elsevier Academic, London).
6. Rohwer F, Edwards R (2002) The phage proteomic tree: A genome-based taxonomy for phage. *J Bacteriol* 184:4529–4535.
7. Simmonds P, et al. (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 15:161–168.
8. Adams MJ, Lefkowitz EJ, King AM, Carstens EB (2013) Recently agreed changes to the International Code of Virus Classification and Nomenclature. *Arch Virol* 158: 2633–2639.
9. Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 106:19126–19131.
10. Peterson AT (2014) Defining viral species: Making taxonomy useful. *Virol J* 11:131.
11. Simmonds P (2015) Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J Gen Virol* 96:1193–1206.
12. Krause DJ, Whitaker RJ (2015) Inferring speciation processes from patterns of natural variation in microbial genomes. *Syst Biol* 64:926–935.
13. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 102:2567–2572.
14. Lopes A, Amarir-Bouhram J, Faure G, Petit MA, Guerois R (2010) Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res* 38:3952–3962.
15. Bobay LM, Touchon M, Rocha EP (2013) Manipulating or superseding host recombination functions: A dilemma that shapes phage evolvability. *PLoS Genet* 9: e1003825.
16. Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 3:199–208.
17. Simmons SL, et al. (2008) Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol* 6:e177.
18. Cadillo-Quiroz H, et al. (2012) Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol* 10:e1001265.
19. Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* 12:263–273.
20. Shapiro BJ, et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48–51.
21. Bobay LM, Ochman H (2017) Biological species are universal across life's domains. *Genome Biol Evol* 9:491–501.
22. Marston MF, Amrich CG (2009) Recombination and microdiversity in coastal marine cyanophages. *Environ Microbiol* 11:2893–2903.
23. Gregory AC, et al. (2016) Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 17:930.
24. Marston MF, Martiny JB (2016) Genomic diversification of marine cyanophages into stable ecotypes. *Environ Microbiol* 18:4240–4253.
25. Cordero OX (2017) Endemic cyanophages and the puzzle of phage-bacteria coevolution. *Environ Microbiol* 19:420–422.
26. Bolduc B, et al. (2017) vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* 5:e3243.
27. Meyer JR, et al. (2016) Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science* 354:1301–1304.
28. Hatfull GF, et al. (2010) Comparative genomic analysis of 60 mycobacteriophage genomes: Genome clustering, gene acquisition, and gene size. *J Mol Biol* 397:119–143.
29. Pope WH, et al.; Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science; Phage Hunters Integrating Research and Education; Mycobacterial Genetics Course (2015) Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife* 4:e06416.
30. Duchêne S, Holmes EC (2018) Estimating evolutionary rates in giant viruses using ancient genomes. *Virus Evol* 4:vey006.
31. Martinsohn JT, Radman M, Petit MA (2008) The lambda red proteins promote efficient recombination between diverged sequences: Implications for bacteriophage genome mosaicism. *PLoS Genet* 4:e1000065.
32. Cain AJ (1954) *Animal Species and Their Evolution* (Hutchinson House, London).
33. Liao CL, Lai MM (1992) RNA recombination in a coronavirus: Recombination between viral genomic RNA and transfected RNA fragments. *J Virol* 66:6117–6124.
34. Faure-Della Corte M, et al. (2010) Variability and recombination of clinical human cytomegalovirus strains from transplantation recipients. *J Clin Virol* 47:161–169.
35. Sijmons S, Van Ranst M, Maes P (2014) Genomic and functional characteristics of human cytomegalovirus revealed by next-generation sequencing. *Viruses* 6:1049–1072.
36. Wilkinson DE, Weller SK (2003) The role of DNA recombination in herpes simplex virus DNA replication. *IUBMB Life* 55:451–458.
37. Nagy M, Nagy E, Tuboly T (2002) Sequence analysis of porcine adenovirus serotype 5 fibre gene: Evidence for recombination. *Virus Genes* 24:181–185.
38. Benko M, Harrach B, Russell WC (2000) Family Adenoviridae. *Virus Taxonomy: Classification and Nomenclature of Viruses*, eds Regenmortel MHVv, et al. (Academic, San Diego).
39. Moss B (1996) Poxviridae: The viruses and their replication. *Fields' Virology*, eds Fields BN, Knipe DM, Howley PM (Lippincott-Raven Publishers, Philadelphia), 3rd Ed, pp 2637–2671.
40. Lima-Mendez G, Toussaint A, Leplae R (2007) Analysis of the phage sequence space: The benefit of structured information. *Virology* 365:241–249.
41. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* 25:762–777.
42. Iranzo J, Koonin EV, Prangishvili D, Krupovic M (2016) Bipartite network analysis of the archaeal virosphere: Evolutionary connections between viruses and capsidless mobile elements. *J Virol* 90:11043–11055.
43. Corel E, Lopez P, Méheust R, Bapteste E (2016) Network-thinking: Graphs to analyze microbial complexity and evolution. *Trends Microbiol* 24:224–237.
44. Iranzo J, Krupovic M, Koonin EV (2017) A network perspective on the virus world. *Commun Integr Biol* 10:e1296614.
45. Aiewsakun P, Simmonds P (2018) The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification. *Microbiome* 6:38.
46. Juhala RJ, et al. (2000) Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 299:27–51.
47. Casjens SR (2008) Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. *Res Microbiol* 159:340–348.
48. Mavrich TN, Hatfull GF (2017) Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol* 2:17112.
49. Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. *Proc Natl Acad Sci USA* 104:2043–2049.
50. Burt A (2000) Perspective: Sex, recombination, and the efficacy of selection–Was Weismann right? *Evolution* 54:337–351.
51. Stebbins GL (1963) Perspectives. I. Animal species and evolution by Ernst Mayr, a review. *Am Sci* 51:362–370.
52. White MJ (1978) *Modes of Speciation* (Freeman, San Francisco), p 456.
53. Ramesh MA, Malik SB, Logsdon JM, Jr (2005) A phylogenomic inventory of meiotic genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Curr Biol* 15:185–191.
54. Malik SB, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM, Jr (2007) An expanded inventory of conserved meiotic genes provides evidence for sex in Trichomonas vaginalis. *PLoS One* 3:e2879.
55. Lahr DJ, Parfrey LW, Mitchell EA, Katz LA, Lara E (2011) The chastity of amoebae: Re-evaluating evidence for sex in amoeboid organisms. *Proc Biol Sci* 278:2081–2090.
56. Cutter AD, Jovelin R, Dey A (2013) Molecular hyperdiversity and evolution in very large populations. *Mol Ecol* 22:2074–2095.
57. Romiguier J, et al. (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515:261–263.
58. Roux C, et al. (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol* 14:e2000234.
59. Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brüssow H (2004) Phage-host interaction: An ecological perspective. *J Bacteriol* 186:3677–3686.
60. Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 8:317–327.
61. Smith GR (1983) General recombination. *Lambda II*, eds Hendrix RW, Roberts JW, Stahl FW, Weisberg RA (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), pp 175–210.
62. Coffin JM (1979) Structure, replication, and recombination of retrovirus genomes: Some unifying hypotheses. *J Gen Virol* 42:1–26.
63. Kim MJ, Kao C (2001) Factors regulating template switch in vitro by viral RNA-dependent RNA polymerases: Implications for RNA-RNA recombination. *Proc Natl Acad Sci USA* 98:4972–4977.
64. Simon-Loriere E, Holmes EC (2011) Why do RNA viruses recombine? *Nat Rev Microbiol* 9:617–626.
65. Kitchen A, Shackelton LA, Holmes EC (2011) Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proc Natl Acad Sci USA* 108:238–243.
66. Geoghegan JL, Duchêne S, Holmes EC (2017) Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog* 13:e1006215.
67. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE (2016) Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 40:258–272.
68. Shi M, Zhang Y-Z, Holmes EC (2018) Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res* 243:83–90.
69. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017) Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* 27: 626–638.
70. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23:673–679.
71. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
72. Bobay LM, Rocha EP, Touchon M (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* 30:737–751.
73. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
74. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30:772–780.
75. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
76. Rambaut A, Grassly NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 13:235–238.
77. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.