

A clash of ideas – the varying uses of the “species” term in virology
and their utility for classifying viruses in metagenomic datasets

Peter Simmonds

*Peter Medawar Building for Pathogen Research, Nuffield Department of Medicine, University of
Oxford, South Parks Road, Oxford, OX1 3SY, UK*

Corresponding author: Peter Simmonds
Peter.Simmonds@ndm.ox.ac.uk
Tel. +44 1865 281 233

Keywords: Species; Taxonomy; Concepts; Virus; Metagenomics; Classification

Word count: Abstract: 250 words
Main text: 6948 words

Non-standard abbreviations:
International Committee for the Taxonomy of Viruses ICTV
High throughput sequencing HTS

ABSTRACT

Species definitions of viruses are frequently descriptive, with assignments often being based on their disease manifestations, host range, geographical distribution and transmission routes. This method of categorising viruses has been recently challenged by technology advances, such as high throughput sequencing. These have dramatically increased knowledge of viral diversity in the wider environment that dwarfs the current catalogue of viruses classified by the International Committee for the Taxonomy of Viruses (ICTV). However, because such viruses are known only from their sequences without phenotypic information, it is unclear how they might be classified consistently with much of the existing taxonomy framework. This difficulty exposes deeper incompatibilities in how species are conceptualised. The original species assignments based on disease or other biological attributes were primarily descriptive, similar to principles used elsewhere in biology for species taxonomies. In contrast, purely sequence-based classifications rely on genetic metrics such as divergence thresholds that include or exclude viruses into individual species categories. These different approaches bring different preconceptions about the nature of a virus species, the former being more easily conceptualised as a category with a part/whole relationship of individuals and species while species defined by divergence thresholds or other genetic metrics are essentially logically defined groups with specific inclusion and exclusion criteria. While descriptive species definitions match our intuitive division of viruses into natural kinds, rules-based genetic classifications are required for viruses known from sequence alone, whose incorporation into the ICTV taxonomy is essential if it is to represent the true diversity of viruses in nature.

MAIN TEXT

The concept of “species”. The classification of animals, plants, fungi and microbes provides a catalogue of the extraordinary diversity of life on the planet. It allows organisms to be named and assigned as species. It also creates higher order taxonomic groupings right up to the levels of kingdom and domain; in our case, from *Homo sapiens* to Metazoa (the kingdom of animals) and Eukaryota (domain) with a further 11 taxonomic layers in between. Of course, applying this man-made system of classification to the biological realm is artificial; clunky man-made categories cannot reproduce the intricate inter-relationships of real biological populations. Nevertheless the division is both functional in grouping related organisms together and provides an evolutionary road map for what ultimately can be described as a “Tree of Life” [1].

Central to biological classification is the concept of species [2, 3]. Typical attributes of species include the ability of members of the same species to inter-breed [4], that they are reproductively isolated from other species and that, collectively, members share a common evolutionary origin and gene pool and that they are subject to natural selection that drives their adaptive fitness and long term evolution [5]. However, this combination of properties is somewhat dependent on what is being studied, since, for example, the inter-breeding requirement cannot apply to asexual organisms, which nevertheless might still form isolated and separate populations that are phenotypically and genotypically distinct from those of other species.

The species division has also been used throughout microbiology and the species description and associated nomenclature of bacteria are key elements of their current taxonomy [6]. The species taxonomic rank is also fundamental to the classification of viruses developed by the International Committee for the Taxonomy of Viruses (ICTV), an organisation that additionally records and

regulates their further assignments into the higher taxonomic ranks of genus, family and, for many virus groups, into orders.

A simple equation between disease and species assignment formed the basis of much of the original descriptive classification of virus species. In the same way that individual species of bacteria may cause specific diseases (tuberculosis is caused by *Mycobacterium tuberculosis*, cholera by *Vibrio cholera*), it has been customary to assign species status to viruses in an equivalent way. Yellow fever is caused by members of the species *Yellow fever virus*, economically devastating disease of tobacco is caused by members of the species, *Tobacco mosaic virus*. There are many hundreds of similar equations in the current virus classification.

However, virologists have recently become considerably more effective at creating datasets of metagenomic data in which viral sequences can be identified simply by random sampling of samples collected from the environment. High throughput sequencing (HTS) allows viral RNA or DNA sequences to be read at considerable depth in any sample type. Application of this technology to seas, lakes, or terrestrial samples has revealed the existence of a staggeringly rich and diverse population of viruses throughout all environments, with perhaps 10-100 times more virus particles than host cells. This equates to around 10^{30} virus particles in existence at any one given time [7, 8]. In these populations, viruses are spectacularly diverse in their complements of genes, genome organisation and types and configurations of genomic nucleic acids.

While basic aspects of the genome organisation, replicative functions of encoded genes, and organisation of structural and accessory genes may be inferred from sequences assembled from metagenomic datasets, the actual viruses they represent are so diverse and unfamiliar that how to place them into the existing ICTV virus classification is problematic. Furthermore, with at best incomplete or most often entirely missing information on virion structure, their host range,

epidemiology, and pathogenicity, we lack the type of phenotypic characteristics that have been used in the past for species descriptions. Parallel advances in bacteriology, mycology and other areas of biology have created similar problems – there are, for example around 13,500 bacterial species formally described, named and with type isolates deposited in international repositories [6], but large scale sequence acquisition of uncultured bacteria suggests there might be many millions of further species in the environment [9, 10]. There is a similar disconnect between named species of unicellular fungi and the many millions predicted from genomic data [11].

This article describe the various ways in which virologists have sought to define species and explore the suitability of these different classification approaches to incorporate viruses found in the much larger datasets of metagenomic sequences obtained from environmental sampling. An appropriate method for assignment of species and higher level taxonomic groupings is essential if we are to follow the recently published Consensus Statement on metagenomic virus classification [12]. This recommends that viruses identified in metagenomic datasets that approximate to complete genome sequences should form part of a greatly extended future ICTV taxonomy. Furthermore, such taxonomic assignments may be made in the absence of direct phenotypic information on the virus. How the ICTV can implement this proposal, however, requires a fundamental re-evaluation of how existing species and other taxonomic rank assignments have been made historically and could be made in the future.

Virus Taxonomy. The classification of viruses (and other micro-organisms) has been extraordinarily helpful in cataloguing viral diversity and making sense of the bewildering array of virion shapes, sizes, genetic content and replication methods. The International Committee for the Nomenclature of Viruses, formed in 1966, was the official body responsible for virus taxonomy and published its 1st Report in 1971 [13]. This listed all classified viruses and their standardised names. Successive publications since then have continued to list an ever expanding number of viruses to the current

day. The last printed (9th) Report from the (renamed) International Committee on the Taxonomy of virus (ICTV) lists viruses classified into 6 orders, 87 families, 349 genera and 2284 species [14]. Over time, the involvement of hundreds of virologists from the Study Groups, sub-committees and the Executive committee and the wider virology community has produced an impressive consensus classification and nomenclature that is increasingly used in the current scientific literature.

This consensus and classification practice has been achieved despite the inherent and biologically unique difficulties associated with virus evolutionary relationships; viruses most likely do not have a single evolutionary origin [15-17] and consequently lack any universal genes from which a shared genetics-based phylogeny could be constructed. Unlike bacterial, fungal, plant, and animal classification, virus taxonomy is simply an assemblage of disconnected units, with many virus families and orders showing no obvious relatedness or evidence of shared evolutionary history with any other group. However much we may learn about virus diversity in the future and however advanced we become at reconstructing gene histories, a Universal Taxonomy of viruses will never be a Universal Evolutionary History of viruses, nor can it be joined as a single entity to the Universal Tree of Life.

Virus species. The ICTV virus classification is a utilitarian, man-made division of virus groups that provide a standardised list of virus taxa, their nomenclature and interrelationships through construction of a multi-layered, hierarchical taxonomy. However, throughout the history of the ICTV, the creation of a virus taxonomy has evoked disproportionate interest and scientific passion, particularly at the species level. There is, for example, disagreement and often heated exchanges about the nature of a virus species that parallel controversies elsewhere in biology [5, 18-22] - is a virus species defined as a group with specific phenotypic and genetic attributes [23, 24] or does the species term for viruses equate simply to a genetically-defined grouping that might correspond to an actual set of replicating organisms in the real world [25, 26]? These discussion parallel uncertainty

elsewhere in biology, where much of the literature about species-level classification revolves around the still unresolved nature of a species and the multiplicity of different methods to assign them (comprehensively reviewed in [5]). Most fundamentally, are species entities created by the human mind as natural categories? – or, following Henning [27], should they be regarded more as real objects, such as evolutionary lineages, that exist in the world independently of humans to conceptualise them (this dichotomy is elegantly discussed in [18]).

Secondly, there is no clear idea about what virus properties might be used to define or describe a species in a way that would be equivalent to the use of the term elsewhere in biology. Viruses are extremely small and effectively intangible, and typically manifested only by their ability to cause disease in their hosts; these restricted observational properties cannot match the precision of the often very detailed phenotypic descriptions of animal and plant species used in for example, morphological or ecological species concepts [28, 29]. Reproductive isolation is a commonly used to delineate species boundaries of animals, plants and fungi in the classical biological species concept [4], but this property cannot be used for virus species delineations because they do not inter-breed in any conventional sense.

The systematic use of the species rank in virus classification commenced in the 7th Report in 1999 and embraced the concept of the species as a class, a group to which viruses may be assigned based on their possession of a common set of properties. A widely used definition first advocated in 1991 is as follows:

“A virus species is a polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche” [30]

In practice, what those properties were, however, could vary considerably between species but were typically based on phenotypic attributes, such as host range, antigenicity, epidemiology and distribution (“ecological niche”). There is also the further requirement that viruses assigned to a species are a single genetic lineage with a common ancestor distinct from all other viruses. The term “polythetic” encompasses the idea that species share a range of properties, typically relating to their epidemiology, host range and pathogenicity, and to viral attributes such as their genome organisation and possession of homologous genes [31, 32].

The current ICTV definition of a species has been recently amended to:

“A species is the lowest taxonomic level in the hierarchy approved by the ICTV. A species is a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria”

This provides a similarly descriptive element to the definition of species but without the requirement that species be polythetic. The ICTV considered that the meaning of that term was both unclear and unrealistic – using polythetic criteria, species may potentially contain members with no common species-defining elements [33]¹. However, for these and other related species definitions used by the ICTV, defining properties of viral species have been primarily descriptive over many decades. The criteria used for assignments have furthermore varied considerably between virus groups, a fact that is immediately evident if one peruses the ICTV 9th Report. Here we learn that bacteriophage species

¹ The polythetic definition has been widely used in biological classification [34] with membership “defined by a combination of characters, each of which may occur also outside the given class and may be absent in any member of the class. . . . Contrary to the situation with universal classes, no single property is either necessary or sufficient for membership in a polythetic class”. The suitability of this species concept and practical applicability for virus classification have however been questioned [33]. As defined, two members of a polythetic species may conceivably share no properties at all that differentiate them from other species, and it is therefore not possible, by definition, to determine whether they “constitute a (distinct) replicating lineage” which represents another element of the virus species definition.

may demarcated by various combinations of host range, tail length of their virions, genome organisation, insertion/deletions, possession of homologous genes and degree of sequence similarity. Animal virus species assignments are often based on attributes of their isolates, such as virion size and morphology, haemagglutination, serological cross-reactivity, genomic attributes such as G+C content, gene complements, nucleotide or amino acid sequence divergence, ability to recombine, and various aspects of their pathology, such as oncogenicity, disease severity and symptomatology.

The other stated element in the virus species definition is that a species is a class into which member viruses may be assigned and thus matches the nature of other ranks in virus taxonomy (genus, family, order). However, this definition is not shared elsewhere in biology, where species concepts create categories that typically possess a part/whole relationship between the individual and the species to which it is assigned [18]. This latter review provides the example of the Earth; this may be classified into the category, “planet”, but planets themselves are actual entities with physical properties such as large size, round shape and general rockiness. Hey then argues that a similar relationship thus exists between an individual polar bear and the biological species to which it is assigned. It’s then a small step to similarly consider a virus strain or isolate to be an example of a virus species that possesses the properties used in its description.

As described above, there are indeed alternative formulations of evolutionary or phylogenetic species concepts that regard species as real biological entities, corresponding to the actual replicating lineage of the species [19, 21, 35]. To take a virological example, the species, *Zika virus* might describe the vast pool of circulating Zika virus particles and replicating entities within the cells and hosts they infect. The species is thus a physical entity made up of swarms of viruses replicating in mosquitoes, birds and a range of mammals. This conception of a viral species can only be described, not defined. Unlike a class with specific inclusion and exclusion criteria, *Zika virus* can

remain as a species even as it evolves and spreads beyond the criteria by which it was originally described - spectacularly in this case following its recent emergence in South America.

The concept of species and categorisation of the natural world. Both evolutionary groups and polythetic or other descriptive definitions of species create entities that match closely to our underlying categorisation of the natural world. This ability is based around category formation, a cognitive process that is central to our ability to organise and extract meaning from sensory information and to use this information. (A fuller description of categories and their expression in language is provided in Supplementary Data.) Concepts are used to categorise objects and actions around us, many of which correspond to individual biological species or genera, such as humans and dogs (members of individual species), oak trees and chimpanzees (genera) and a variety of broader categories such as fish and trees that match up approximately but often imprecisely with taxa higher up the Linnean classification. As a result, there is a considerable overlap between our built-in conceptualisation of individual species of animals or plants and their formal biological definition.

While seemingly a reasoned process, our ability to create categories is a hard-wired, largely unconscious process. Furthermore, although we typically have words for the various categories we form, an ability to form categories is not dependent on language and developed very early in vertebrate evolution [36, 37]. There is clear evidence from experimental studies that birds and non-human mammals categorise the world very much as we do [36]; without a language ability they are simply unable to convey that to us [25]! As a result, the mapping of a real world object to its spoken or written linguistic expression is indirect. This is exemplified by relationships between adenoviruses in the real world, a virologist's conception of an adenovirus virion and its linked associations, and the way in which this concept can be expressed in language as a token that activates the same or a similar concept in a reader or listener (Fig. 1). The recognition of such virus particles by a knowledgeable virologist will bring to the fore what is essentially an idealised concept of an

adenovirus, perhaps primarily an internal image of the virus, that may further evoke a visualisation of its genome organisation and splicing patterns, or for a clinician, images of the various disease manifestations of adenovirus infections (top of triangle). Any of these internal concepts can be expressed as “adenovirus” in communication with others (left, bottom triangle), although the extent to which a writer’s and reader’s conceptions of adenovirus overlap can be highly variable and context dependent. Basic and clinical virologists often have difficulties in communicating even when talking about the same virus because their internal concepts of the virus are so different.

The important aspect of this semiotic triangle is that codification of objects in language only functions if there is an internal categorisation of the object that can map one side of the triangle to the other. The word “adenovirus” does not directly map to the real world object. We might imagine that “words mean things”, but the direct link between the object in the real world and its written or spoken name is simply not there [25, 38].

The relevance of the use of concepts and their codification in language lies in the way that this internal process profoundly influences the way in which we approach the biological world and the way in which we seek to classify groups within it. This is all the more important since we are largely unaware of how internal categorisation takes place despite possessing strong intuitions about what seems right and wrong when we try to classify things. One of the most surprising attributes about category formation is that species defined in this way are not logically defined entities with specific inclusion and exclusion criteria. Indeed, the associated requirement that categories are mutually exclusive and that each entity must belong unequivocally to one, and only one, of the proposed categories provides a wholly unrealistic basis for the categorisation we use in everyday life (see Supplementary Data). Necessary and sufficient conditions are rarely encountered and defining properties of a category are frequently continuous variables without clear cut boundaries.

Cognitively, categories are actually more usually formed using prototypes [39, 40], which possess the idealised properties of the category (rather like a Platonic *Form*) and with members sharing variable numbers of its attributes (a “radial” structure). Those closest to the prototype can be considered as better examples of the category than those sharing fewer attributes, thus a herring may better represent the category “fish” than a sea-horse [39, 40]. An alternative formulation is the exemplar, in which categories may be internally represented as a series of examples rather than an idealised prototype [41]. A perceived object may thus be compared to multiple known exemplars in a category for identification. Exemplars can make better sense of some categories such as Wittgenstein’s well-known example of “games”, a polythetic category for which there is ultimately no commonality in all its various forms [42](see Supplementary Data).

Are natural categories the best way to classify virus species? Historically, the classification of viruses has followed practices elsewhere in biology, where species might be assigned using descriptions of their biological properties (pathogenicity, epidemiology, morphology), their resemblance to prototypes and their associated polythetic descriptive species definitions. This arguably more intuitive approach contrasts markedly with species definitions that are based upon defined necessary and sufficient inclusion and exclusion criteria that define a logical class. The use of cognitive categories indeed substantially shapes our concepts of species and the usage of the term in virology, often quite subconsciously - seemingly simple descriptive definitions of viral species bring with them all sorts of cognitive attributes and baggage that accompanies category formation in other areas of our internal representation of the world. A premise of this review is that these pre-conceptions and assumptions are at the source of ongoing difficulty and academic controversy with those espousing largely genetics-based classification approaches. These can be summarised as follows.

Defining virus species. The idea that categories are formed through the use of prototypes or exemplars matches closely with polythetic or other descriptive definitions of viral species. Members of virus species may typically possess a series of properties that are contained within the specific definition / description, but without any of them being absolutely required. Category formation entails the use of virus species definitions that are intrinsically descriptive, each requiring a range of properties on which to base a definition, or in cognitive terms, around a prototype. If we return to the ICTV Report and review the range of species definitions provided in the various families, we do indeed find that the majority of species definitions are coined in this way and typically lack the precisely defined inclusion and exclusion criteria expected of a logical class. As an example, the large number of *Flavivirus* species are typically described in terms of their geographical location, mode of transmission (*ie.* vectored by which mosquito species) and likelihood of causing disease, but without any indication of which of these attributes are necessary or dispensable for species membership.

Both descriptive and evolutionarily-based species assignments create a direct mapping of the species to categories, natural kinds or actual objects, such as evolutionary lineages. And in many ways, “species as objects” can represent an alternative way to describe a virus species. As an example, flaviviruses have frequently emerged into new geographical locations, they can change vectors and their pathogenicity can be variable in different populations. And their exact properties are often contingent on incidental historical events – a mosquito infected with West Nile virus carried in a flight from Israel to New York in 1999 brought devastation to bird populations throughout Northern America and many, sometimes fatal human infections [43]. Thus the geographical range element of the species description *West Nile virus* before then was evidently not, as it turned out, a necessary or defining property of the species, it was just an incidental restriction of its vector distribution. In the many descriptive definitions of virus species there similarly lurk simple historical or epidemiological events that govern the behaviour and range of the virus and are not intrinsic properties of the virus species itself. *West Nile virus* did not stop being *West Nile virus*

just because it spread to North America, even though its extended distribution fell outside its original description or definition.

Virus and species names. Elsewhere in biology, the use of descriptive species definitions conceptually underpins the widespread use of species names to directly refer to the organisms they classify in addition to the names of the class to which these real world objects are assigned. Species names that refer to real world objects are indeed required where common names are lacking or so variable in different languages that the Linnaean designation is preferable as a *lingua franca*. These includes such commonly described organisms as *Escherichia coli* (a bacterium) and *Drosophila melanogaster* (a type of fruit fly) where names have not been coined in any language. Indeed, this extends to the vast majority of the several million classified plants, fungi, and bacteria which similarly lack common names.

The part/whole equation of species to the objects they describe and the consequent blurring of virus and species names [44] is however incompatible with current ICTV taxonomy rules that specify that a virus species is a class not an object [24]. Insistence of this typological convention that is largely specific to virology creates a whole range of problems and confusion with terminological usage that simply do not exist in other areas of biological classification. Indeed, the strict use of the “species as class” definition often requires that virus taxonomy maintains parallel (and often somewhat redundant) nomenclature for labelling the objects such as West Nile virus from the species into which it is classified - *West Nile virus* in the above example. The difference here is the italicisation and certain restrictions in the orthography of the species name. The consequent rigid restrictions on virus taxon names means that sentences such as “Infections with *Enterovirus A* may be associated with neonatal encephalitis” are incorrect, ostensibly because patients cannot be infected with a taxonomic category, only by a virus. However, they might quite reasonably be infected with *Escherichia coli*! Perhaps, as has been suggested previously [26, 44], the ICTV might re-consider the

rigid nomenclatural distinction between species as a class and the virus populations that are assigned to it. A relaxation of the distinction where the context makes it obvious whether a virus or a virus taxon was being referred to would bring virology very much more into line with microbiological usage, and with use of the species term throughout the rest of biology.

The imposition of parallel nomenclature for viruses and the species to which they are assigned and the status of species as a class is historically relatively recent in virology. Indeed, plant virologists, with their plethora of entirely descriptive species names have been until recently, resistant to formally adopt species rank in their classifications [45] and maintained the designation “virus” in their descriptions of the various groups of plant viruses. While species names were ultimately adopted in the 1990s, they are even now, with a few exceptions nothing more than an italicised version of the original virus name, *ie.* banana bunchy top virus simply becomes *Banana bunchy top virus*. Largely comprised of single members, species are still largely alternative formulations the names of the viruses, and this nomenclatural practice remains wedded to the descriptive type of species definition. Such usage does not really embrace the intended status of a species as a class but it does reflect how the individual plant virus species are conceptualised in practice.

What should be classified? Inherited from cognitive categories is the dependence on a rich and informative set of attributes to describe a virus species. There is consequently a widely expressed opinion that only such viruses possessing such properties are worthy subjects of classification [23]. Indeed, much of the information used historically to describe virus species has been based upon their pathogenicity and epidemiological properties, as well as their genome organisation and genetic relationships to members of other taxa. This approach to classification cannot work, however, for viruses where we don’t have any phenotypic properties to start with, as with the tens or hundreds of thousands of viral sequences derived from environmental samples. They may also be entirely bereft of any particular genomic features other than genetic divergence from other viruses to define them.

As a community, we have found it difficult (or perhaps counter-intuitive, pointless or arbitrary) to place these into the ICTV classification using the existing taxonomy proposal system, which is better suited for assignment criteria beyond simple metrics of genetic divergence. I argue that at least part of this difficulty originates because we have become accustomed to using much richer information sets that enables virus descriptions and properties to be conceptualised internally. Without this, we would be reduced to remembering random names, numbers or codes in designating virus species assigned using minimal genetic criteria. Even though this may be computationally trivial, humans are hopeless at this type of task, as almost all 7 year olds learning the times tables will tell you!

Genetic classification of viruses. The challenge of classifying viruses derived from metagenomics datasets arises from the lack of phenotypic data that has been used historically in the largely descriptive species definitions of currently classified viruses. Indeed, the early classification of viruses occurred when sequencing methods were in their infancy and the nucleotide sequences of most viruses were unknown. Confronted as we now are with vast numbers of viruses in metagenomics datasets, can new species be assigned by sequence data alone and how would such species descriptions or definitions match those in the current ICTV classification in practical and conceptual terms?

To explore this issue, we can first put the question the other way around - could we derive a satisfactory virus classification using sequence data alone from currently assigned viruses in the existing ICTV taxonomy? We do now have comprehensive, complete genome sequence data available for members of almost all of the 3000+ species in the current ICTV taxonomy. In general, there is an encouraging although not perfect match between their sequence relationships, such as gene phylogenies and their species assignments based on phenotypic properties. Furthermore,

members of almost all species form single evolutionary lineages required by the current ICTV species definition².

However, as might be expected, there is a highly variable relationship between genetic divergence and their phenotypic properties. For example, members of different flavivirus species show pairwise distances in the region of NS5 used for family-wide comparisons ranging from as little as 1.5% (between the species *Israel turkey meningoencephalomyelitis virus* and *Bagaza virus*) to a maximum value of 44% (*San Perlita virus* and *Ilheus virus*). Distances between members of the four assigned species in the *Pestivirus* genus range from 13%-17% but in this case, one of the original species, *Bovine viral diarrhoea virus* had to be split into two species based on sequence divergence between members. Similarly variable relationships between sequence divergence and classification are found widely between other virus families. In some families, different species are highly similar genetically, while, for example, those in the genus *Mitovirus* in the yeast virus family *Narnaviridae* show 64%-82% divergence and their sequences can barely be aligned.

Despite the variability between sequence divergence and species assignments in different virus groups, the ready availability of nucleotide sequences for newly characterised viruses has led to an increasing tendency to base species description and/or definitions on measures of sequence relatedness. In some cases, sequence divergence may become the principal assignment criterion. For example, several new species of hepatitis E virus (*Orthohepevirus B-D*) have recently been assigned on the basis of a threshold of 30% amino acid sequence divergence in conserved domains of methyltransferase, helicase and polymerase genes [46]. Genotypes within these species have been assigned using similar, lower divergence thresholds. Such divisions have been made without requiring any knowledge of the abilities of members of the various taxonomic groups formed to

² There are a few exceptions, including the grouping of members of the flavivirus species *Louping ill virus* within the members of another species, *Tick-borne encephalitis virus*.

cause disease or of their cellular tropisms. Or indeed, virological properties - for most taxa, only assembled virus sequences have been obtained, not virus isolates.

In this example, the classification of orthohepeviruses by genetic relatedness does not naturally follow criteria that might have used previously to classify them, such as host range and epidemiology. For example, members of *Orthohepevirus A* can infect humans, deer, mongooses, pigs, wild boar, and camels, while the other orthohepevirus species (B-D) are restricted in host range to birds, rodents, and bats respectively. The range of hosts known to be targeted by different hepeviruses may indeed change as more mammalian species are sampled in the future, as indeed may the number of species assigned. The important point is that sequence relatedness is the primary means of assignment and it divides viruses in this family into groups that may only incidentally relate to their phenotypic properties.

Although not labelled differently in the ICTV taxonomy, the species assignments made on the basis of a single property, genetic divergence, represents a quite different type of species definition from the phenotypically-driven descriptive approaches described in the previous section. Such species therefore make no use, declared or otherwise, of the prototype or polythetic categories, and may be more likened to phenetic (distance-based) or phylogenetic (lineage-based) species concepts [19, 35, 47, 48]. They differ too in their reliance on objective criteria, such as sequence divergence in a defined genome region; these represent the type of necessary and sufficient conditions of Aristotelian classification and the creation of classes with strictly delineated membership rules. Furthermore, genetic classification makes it considerably easier to conceptualise species as classes since the inclusion or exclusion of a virus as a member of a species can be based upon a simple objective criterion.

Another difference from descriptive species definitions is that the species rank has no special status, either as a defined degree of sequence divergence or position in a phylogeny. In contrast to earlier formulated phylogenetic specific concepts of species, such as “ ... *the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent*” [49], the existence of virus strains and below species classification of genotypes or perhaps subspecies mean that viral species often do not form the lowest grouping in a phylogenetic tree. The apparent arbitrariness of the species and higher taxon levels has led some to advocate entirely rank-free classifications of organisms, such as the Phylocode (<https://www.ohio.edu/phylocode/>), where all named taxonomic levels including species are expunged [19].

These approaches have already caused considerable discomfort among virologists who have previously advocated the polythetic and descriptive criteria for taxonomic assignments. The following text from Marc van Regenmortel [23] expresses this view very cogently:

The proposers of the new definition (King 2012) also dismissed the glaring case of the 288 begomovirus species that were created by ignoring the polythetic principle and accepting that species could be established on the basis of a single arbitrary criterion, namely less than 89 % pairwise sequence identity in the viral DNA-A genome (Fauquet et al. 2003). Many of the so-called 288 different ‘species’ consist of viruses that infect the same host (cotton or tomato) and produce very similar disease symptoms, and had to be given different names by including the geographical location of the first isolation of the virus. This produced a long list of species names such as *Tomato leaf curl Comoros virus*, *Tomato leaf curl Guangxi virus*, *Tomato leaf curl Hsinchu virus*, *Tomato leaf curl New Delhi*, etc., which could have been considered strains of the same species if a lower threshold demarcation percentage for creating species had been chosen (Van Regenmortel 2011).

The author of the above text is the originator of the polythetic principle for species definitions and the pre-eminent authority on the typology of taxonomic classes (and ex-president of the ICTV). The passage typifies very clearly the disquiet experienced by many virologists with genetics-based species definitions. These include what can appear as arbitrary choices of divergence thresholds used to delineate species, the authors' perception of the meaningless status of taxa then created and the pointless and uninformative nomenclature that is then required for them.

I believe the (unstated) root of the problem is the desire for species divisions to “mean” something and which group viruses according to tangible and distinctive characteristics that can be internally conceptualised. Returning to the semiotic triangle (Fig. 1), this type of classification approach is more a case of using rules to codify objects in the world without engaging our remarkable and hard-wired abilities to first create internal categories. Traversing the base of the triangle, as this process does, can seem intuitively “wrong” even though, as we shall see, internal concepts and natural categorisation through virus descriptions may become unnecessary luxuries as the pace of virus discovery accelerates.

Classifying metagenomic sequences into viral taxa. With this background, we may be better able to consider the challenges presented by classifying the vast number of viruses from sequences generated by HTS from environmental and biological sampling [7, 50-52]. Incorporating these into the ICTV taxonomy requires a number of assumptions and changes to current taxonomic practice. First, a decision to classify metagenomic data tacitly acknowledges that a virus sequence in an HTS dataset represents a virus as real as conventional viruses whose replication properties, pathogenicity, and epidemiological attributes can be used for their classification. Indeed, a virus represented in the HTS dataset possesses an equally rich set of properties; it's simply that we have not had the opportunity to find out what these are. I believe that most practicing virologists agree with this assumption.

505

506 A second factor motivating the classification of viruses from metagenomic data is the general
507 agreement that that an important purpose of virus taxonomy is to catalogue the true diversity of
508 viruses in the world and not to place arbitrary and utilitarian limits on what is “worth” or “not
509 worth” classifying. This was the unanimous view of the recently convened expert group that
510 produced the Consensus Statement on classifying metagenomic sequences [12] and a view that is
511 widely shared elsewhere in virology. There are indeed similarly motivated classification bodies in
512 other areas of microbiology that have to deal with comparable disparities in the numbers of formally
513 named bacterial and fungal species and the undoubted existence of perhaps several million such
514 species in nature [9, 11].

515

516 At a technical level, there is a further requirement that sequences of viruses in metagenomic
517 datasets are free of the numerous potential errors and artefacts that originate from HTS methods
518 and environmental sampling. Particularly problematic is the risk of assembling artificially chimaeric
519 genomes from mixed virus populations within a sample. Evidence is further required that assembled
520 sequences are complete (or at least coding complete) as required by the ICTV [53], enabling a more
521 complete genetic characterisation of this virus and its relationships with other taxa. This
522 requirement is particularly problematic for the recovery of complete genome sequences of viruses
523 with segmented or multipartite genomes that need to be plausibly linked together. For entirely
524 novel taxa, it may be unclear at the outset how many segments actually represent a complete virus
525 genome. A different problem lies in the differentiation of exogenous viruses from those integrated
526 into host genomes through earlier endogenization events [54]. These are not fundamental
527 restrictions on the use the metagenomic sequence data and indeed, newer technologies that
528 generate longer reads and template circularisation can circumvent many of the problems of
529 inaccurate assembly and read errors [55-57].

530

Given that virus sequences in metagenomic datasets represent real viruses, that they accurately represent all or almost all of a virus genome, and there is agreement that they should be classified, proceeding with this in a consistent and inclusive way requires methodologies for species and other taxonomic rank assignments that are largely or entirely genetically based even though the actual viruses as physical objects remain elusive and their properties are largely unknown. Such methods would have to be shown to be robust, reliable and to be able to reproduce the division of viruses into orders and families and lower taxa equivalently to how conventional viruses currently are. Given that such methods can be developed, a hypothetical classification pipeline might take HTS sequence data as its raw input and regurgitate vast lists of assignments of the viruses they contain into existing and a likely huge number of newly generated virus taxa. Even with current day more limited genetics analysis methods, any process of purely genetically-based assignments can be represented operationally as traversing the base of the semiotic triangle (Fig. 1), since we are effectively taking real world objects, in the form of HTS data and generating from this a coded output of named or otherwise labelled viruses into new taxonomic categories. This process thus bypasses the cognitive processing and category formation of traditional descriptive taxonomy.

Despite this, the process of algorithmic coding produces labels by which these real-world objects can be referred to and assigns them into groups for which further information may be garnered in the future. Throughout microbiology, there are indeed a multitude of such technically labelled groups, such as *Escherichia coli* O157:H7, which have been propelled into popular consciousness through their recent dramatic emergence and devastating pathogenicity. Returning to the previous example of hepeviruses, their genetic division into the four species provides a useful set of reference groups for future studies of the host distributions, relative pathogenicities, and tissue tropisms (although I doubt only the most hardened plant virologist would ever become conceptually acquainted with the 288 proposed species of begomoviruses!!). The value of such algorithmically-based classifications and their acceptance by the virology community more generally will depend on the extent to which

our initial discomfort with the process can be reconciled with the broader necessity to get such viruses incorporated into the ICTV classification and assignment of labels by which such groups can be referred to and explored in the future. We will also have to sacrifice the privileged status of the species taxon layer elsewhere in biology and to break its association with the actual viral entities we are classifying. Understanding the procedural differences between sequence-based and traditional, descriptive virus classification methods is an important first step.

Conclusions. The article seeks to illuminate the ways in which the virus species, seemingly a single taxonomic category, can actually be conceptualised in quite different ways. Understanding the often unstated principles and assumptions behind species definitions represent important steps in reconciling areas of current and past controversy that surround the subject. To summarise, the assignment of viruses to species and other taxonomic classes in the current ICTV taxonomy can be entirely descriptive or may be based on formal inclusion or exclusion principles that define a logical class (contrasting examples are listed in Table 1). Species definitions listed on the left of the table correspond to the virological equivalents of those used widely elsewhere in biology, being based primarily on distinctive clinical or epidemiological features. These differ markedly from sequence-based or sequence-assisted taxonomy typically based upon sequence divergence thresholds or phylogeny groupings (right hand column).

To be clear, neither descriptive nor entirely sequence-based methods are “right” nor “wrong”, but they differ in their assumptions, typology and relationships to usage elsewhere in biology. Descriptive species definitions may suit best our intuitive division of viruses into natural kinds, but this may entail a series of unstated assumptions and biases that are inappropriate for rule-based genetic classifications. Furthermore, with little or no phenotypic data, viral sequences in metagenomic datasets, although representing viruses every bit as real as those that have been isolated and conventionally characterised, have to be classified using genome sequences alone.

583 While not initially providing the descriptive elements of more traditionally described virus groups,
584 sequence-based assignments can produce a relevant species and higher taxonomy level
585 classification framework that will guide future genetic and phenotypic investigation of their
586 properties.

587

588

589

590

591

592

593

594 FUNDING INFORMATION

595 This work was funded in part by the Wellcome Trust (WT108418AIA)

596

597

598 ACKNOWLEDGEMENTS

599 I would like to thank Mike Adams, Heli Harvala and Donald Smith for thoughtful review of the

600 manuscript and their useful input into creating a more readable and relevant article. The author is a

601 member of the ICTV Executive Committee, but the views expressed in the article are those of the

602 author alone and are not to be equated with those of the ICTV or of any of its other members.

603

604 CONFLICTS OF INTEREST

605 The author declares no conflicts of interest

606

607

608

609

610

611

REFERENCES

1. **Woese CR.** Bacterial evolution. *Microbiological reviews* 1987;51(2):221-271.
2. **Cain AJ.** Linnaeus's *Ordines naturales*. *Archives of Natural History* 1993;20:405-415.
3. **Darwin C.** *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray; 1859.
4. **Mayr E.** *Systematics and The origin of species from the viewpoint of a zoologist*. New York: Columbia University Press; 1942.
5. **Mayden RL.** A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF, Dawah HA, Wildon MR (editors). *Species: the Units of Biodiversity*. London: Chapman and Hall; 1997. pp. 381-424.
6. **Parker CT, Tindall BJ, Garrity GM.** International Code of Nomenclature of Prokaryotes. *International journal of systematic and evolutionary microbiology* 2015.
7. **Edwards RA, Rohwer F.** Viral metagenomics. *Nat Rev Microbiol* 2005;3(6):504-510.
8. **Suttle CA.** Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* 2007;5(10):801-812.
9. **Konstantinidis KT, Rossello-Mora R.** Classifying the uncultivated microbial majority: A place for metagenomic data in the Candidatus proposal. *Systematic and applied microbiology* 2015;38(4):223-230.
10. **Hedlund BP, Dodsworth JA, Staley JT.** The changing landscape of microbial biodiversity exploration and its implications for systematics. *Systematic and applied microbiology* 2015;38(4):231-236.

- 634 11. **Hibbett DS, Taylor JW.** Fungal systematics: is a new age of enlightenment at hand? *Nat Rev*
635 *Microbiol* 2013;11(2):129-133.

- 636 12. **Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR et al.** Consensus statement: Virus
637 taxonomy in the age of metagenomics. *Nat Rev Microbiol*, Consensus Statement
638 2017;15(3):161-168.

- 639 13. **Wildy P.** Classification and nomenclature of viruses. First report of the International
640 Committee on Nomenclature of Viruses. *Monog virol* 1971;5:1-81.

- 641 14. *Ninth Report of the International Committee on Taxonomy of Viruses.* London: Academic
642 Press; 2009.

- 643 15. **Durzynska J, Gozdicka-Jozefiak A.** Viruses and cells intertwined since the dawn of evolution.
644 *Virol J* 2015;12:169.

- 645 16. **Koonin EV, Senkevich TG, Dolja VV.** The ancient Virus World and evolution of cells. *Biology*
646 *direct* 2006;1:29.

- 647 17. **Brussow H.** The not so universal tree of life or the place of viruses in the living world.
648 *Philosophical transactions of the Royal Society of London Series B, Biological sciences*
649 2009;364(1527):2263-2274.

- 650 18. **Hey J.** The mind of the species problem. *Trends in Ecology & Evolution* 2001;16(7):326-329.

- 651 19. **Mishler BD.** Species Are Not Uniquely Real Biological Entities. In: Ayala FJA, R. (editor).
652 *Contemporary Debates in Philosophy of Biology*: Blackwell; 2010. pp. 110-122.

- 653 20. **Claridge RF.** Species Are Real Biological Entities. In: Ayala FJA, R. (editor). *Contemporary*
654 *Debates in Philosophy of Biology*: Blackwell; 2010. pp. 91-109.

- 655 21. **Ghiselin MT.** A radical solution to the species problem. *Syst Zool* 1975;23:536-544.

- 656 22. **Hull DL.** Are species really individuals? *Syst Zool* 1976;25:174-191.
- 657 23. **Van Regenmortel MHV.** Classes, taxa and categories in a heirarchical virus classification: a
658 review of current debates of definitions and names of species. *Bionomia* 2016;(in press).
- 659 24. **Van Regenmortel MH.** Viruses are real, virus species are man-made, taxonomic constructions.
660 *Arch Virol* 2003;148(12):2481-2488.
- 661 25. **Hurford JR.** Animals approach human cognition. *The origins of meaning*. Oxford: Oxford
662 University Press; 2007. pp. 20-64.
- 663 26. **Bos L.** Virus nomenclature; continuing topicality. *Arch Virol* 2003;148(6):1235-1246.
- 664 27. **Hennig W.** *Phylogenetic systematics*. Champaign/Urbana, IL:: University of Illinois Press; 1966.
- 665 28. **Stuessy TF.** *Plant Taxonomy. The Systematic Evaluation of Comparative Data*. New York:
666 Columbia University Press; 1990.
- 667 29. **Cronquist A.** Once again, what is a species? In: Knutson LV (editor). *BioSystematics in*
668 *Agriculture* Montclair, New Jersey: Allenheld Osmun; 1988. pp. 3-20.
- 669 30. **Van Regenmortel MH, Maniloff J, Calisher C.** The concept of virus species. *Arch Virol*
670 1991;120(3-4):313-314.
- 671 31. **Van Regenmortel MH.** Applying the species concept to plant viruses. *Arch Virol* 1989;104(1-
672 2):1-17.
- 673 32. **Beckner L.** *The Biological Way of Thought*. New York: Columnia University Press; 1959. p. 55-
674 80.
- 675 33. **Giibs AJ, Gibbs MJ.** A broader definition of ‘the virus species’. *Arch Virol* 2006;151:1419–1422.
- 676 34. **Beckner N.** *The biological way of thought*. New York: Columbia University Press; 1959.

- 677 35. **Wiley E.** The evolutionary species concept reconsidered. *Systematic Zoology* 1978;27:17-26.
- 678 36. **Vauclair J.** *Animal cognition: An introduction to modern comparative psychology*. London:
- 679 Harvard University Press; 1996.
- 680 37. **Hauser MD.** *Wild Minds: What animals really think*. New York: Henty Holt; 2000.
- 681 38. **Jackendoff R.** *Reference and truth*. Foundations of language. New York: Oxford University
- 682 Press; 1999. p. 294-322.
- 683 39. **Rosch EH.** Natural categories. *Cognitive Psychology* 1973;4:328-350.
- 684 40. **Lakoff G.** *Women, fire and dangerous things: What categories reveal about the mind*. Chicago:
- 685 University of Chicago Press; 1987.
- 686 41. **Nosofsky RM.** Generalized Context Model: An Exemplar Model of Classification. Formal
- 687 Approaches to Categorization. In: Pothos EM, Wills AJ (editors). *Formal approaches in*
- 688 *classification*. Cambridge: Cambridge Universtiy Press; 2011. pp. 18-39.
- 689 42. **Wittgenstein L.** Philosophische Untersuchungen. In: Hacker PMS, Schulte J (editors).
- 690 *Philosophical Investigations, tthe German text with an English translation*. Oxford: Blackwell;
- 691 2009. pp. 1-181.
- 692 43. **Roehrig JT, Layton M, Smith P, Campbell GL, Nasci R et al.** The emergence of West Nile virus
- 693 in North America: ecology, epidemiology, and surveillance. *CurrTopMicrobiolImmunol*
- 694 2002;267:223-240.
- 695 44. **Bos L.** Coming to grips with the naming of viruses; continuing discord, or a way out? *Arch Virol*
- 696 2007;152(3):649-653.
- 697 45. **Martelli GP.** Classification and Nomenclature of Plant Viruses: State of the Art *Plant Disease*
- 698 1992;76:436-442.

- 699 46. **Smith DB, Simmonds P, I, Jameel S, Emerson SU et al.** Consensus proposals for classification
700 of the family Hepeviridae. *J Gen Virol* 2014;95(Pt 10):2223-2232.
- 701 47. **Sneath PHA.** Phenetic taxonomy at the species level and above. *Taxon* 1976;25:437-450.
- 702 48. **de Queiroz K.** The general lineage concept of species and the defining properties of the
703 species category. In: Wilson RA (editor). *Species*. Massachusetts: MIT Press; 1999. pp. 49-89.
- 704 49. **Tanaka K, Lapointe R, Barney WE, Makkay AM, Stoltz D et al.** Shared and species-specific
705 features among ichnovirus genomes. *Virology* 2007;363(1):26-35.
- 706 50. **Rosario K, Breitbart M.** Exploring the viral world through metagenomics. *Current opinion in*
707 *virology* 2011;1(4):289-297.
- 708 51. **Roossinck MJ.** Plant virus metagenomics: biodiversity and ecology. *Annu Rev Genet*
709 2012;46:359-369.
- 710 52. **Simmonds P.** Methods for virus classification and the challenge of incorporating metagenomic
711 sequence data. *J Gen Virol* 2015;96(Pt 6):1193-1206.
- 712 53. **Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL et al.** 50 years of the International
713 Committee on Taxonomy of Viruses: progress and prospects. *Arch Virol*, journal article 2017:1-
714 6.
- 715 54. **Katzourakis A, Gifford RJ.** Endogenous viral elements in animal genomes. *PLoS Genet*
716 2010;6(11):e1001191.
- 717 55. **Acevedo A, Andino R.** Library preparation for highly accurate population sequencing of RNA
718 viruses. *Nature protocols* 2014;9(7):1760-1769.
- 719 56. **Whitfield ZJ, Andino R.** Characterization of Viral Populations by Using Circular Sequencing. *J*
720 *Virol* 2016;90(20):8950-8953.

- 721 57. **Kumar A, Murthy S, Kapoor A.** Evolution of selective-sequencing approaches for virus
722 discovery and virome analysis. *Virus Res* 2017;239:172-179.

723

724

TABLE 1

DESCRIPTIVE AND SEQUENCE-BASED CLASSIFICATION OF VIRUSES

<i>Examples of species with descriptive definitions – typically one virus member per species</i>	<i>Sequence-based species assignment examples - typically multiple members per species</i>
<p><i>Flavivirus</i> genus, family <i>Flaviviridae</i></p> <p><i>Alphavirus</i> genus, family <i>Togaviridae</i></p> <p>Species in the plant virus families: <i>Benyviridae, Bromoviridae, Caulimoviridae, Closteroviridae, Geminiviridae, Iridoviridae, Luteoviridae, Nanoviridae, Partitiviridae, Potyviridae, Secoviridae, Tombusviridae, Tymoviridae, Virgaviridae,</i></p> <p>And many others...</p>	<p><i>Enterovirus</i> genus, family <i>Picornaviridae</i>: 35% amino acid sequence divergence in P1 and 25% in non-structural genes.</p> <p><i>Hepacivirus</i> genus, family <i>Flaviviridae</i>: 25% amino acid sequences divergence in NS3 and 30% divergence in NS5B</p> <p><i>Alpha-, Beta-, Gammatorquetenovirus</i> genera, family <i>Anelloviridae</i>: 35% nucleotide sequence divergence in ORF1</p> <p><i>Lyssavirus</i> genus, family <i>Rhabdoviridae</i>: 18-20% sequence divergence in N gene</p>

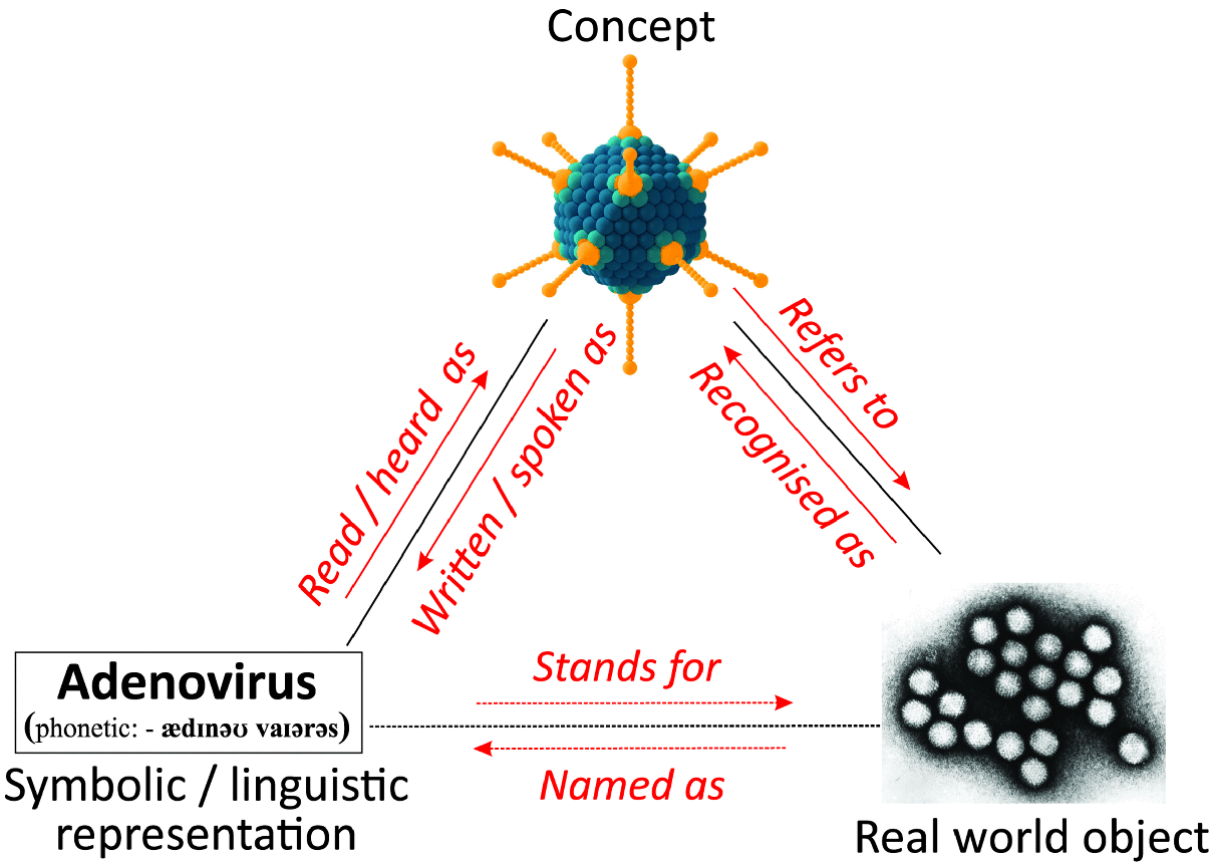


Fig. 1. Conceptualising and naming adenovirus virions: relationships between the real world object, concepts and names. The triangle depicts the relationships between objects in the real world and their internal conceptual and symbolic (eg. named) representations. A virologist’s conception of an adenovirus, a dsDNA virus with a striking virion morphology, is an illustrative example. Adenovirus virions (as real world objects) can be visualised by electron-microscopy and but are conceptualised internally, typically as an idealised representation of the virus particle and its linked associations with replication mechanisms, genome organisation and disease associations (collectively, the concept of adenovirus). Once this category is formed and its attributes learned, its written or spoken linguistic expression as “adenovirus” serves as a pre-loaded token in communication with other virologists where a similar already formed concept may be elicited (at least in those virologists familiar with adenoviruses). While the existence of perception and expression pathways are established, the connections across the base of the triangle are not. Naming objects without

745 establishing an internal conceptual representation first (right to left) or equating words to objects
746 directly (left to right) may not occur in natural cognition or language despite the apparent simplicity
747 of the relationships depicted. A more general account of the semiotic triangle and its cognitive basis
748 is depicted in Fig. S1 (Suppl. Data). Images were obtained from
749 https://commons.wikimedia.org/wiki/File:Adenovirus_3D_schematic.png
750 http://phil.cdc.gov/PHIL/Images/08101998/00042/B82-0142_lores.jpg

751

752

753

754

755