

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/20067> holds various files of this Leiden University dissertation.

Author: Lauber, Chris

Title: On the evolution of genetic diversity in RNA virus species : uncovering barriers to genetic divergence and gene length in picorna- and nidoviruses

Date: 2012-10-30

CHAPTER 3

Toward Genetics-Based Virus Taxonomy:
Comparative Analysis of a Genetics-Based
Classification and the Taxonomy of
Picornaviruses

Chris Lauber
Alexander E. Gorbalenya

Journal of Virology (2012) 86:3905
(JVI **spotlight** feature)

Abstract

Virus taxonomy has received little attention from the research community despite its broad relevance. In an accompanying paper²⁸², we have introduced a quantitative approach to hierarchically classify viruses of a family using pairwise evolutionary distances (PEDs) as a measure of genetic divergence. When applied to the six most conserved proteins of the *Picornaviridae*, it clustered 1,234 genome sequences in groups at three hierarchical levels (to which we refer as the “GENETIC classification”). In this study, we compare the GENETIC classification with the expert-based picornavirus taxonomy and outline differences in the underlying frameworks regarding the relation of virus groups and genetic diversity that represent, respectively, the structure and content of a classification. To facilitate the analysis, we introduce two novel diagrams. The first connects the genetic diversity of taxa to both the PED distribution and the phylogeny of picornaviruses. The second depicts a classification and the accommodated genetic diversity in a standardized manner. Generally, we found striking agreement between the two classifications on species and genus taxa. A few disagreements concern the species Human rhinovirus A and Human rhinovirus C and the genus Aphthovirus, which were split in the GENETIC classification. Furthermore, we propose a new supergenus level and universal, level-specific PED thresholds, not reached yet by many taxa. Since the species threshold is approached mostly by taxa with large sampling sizes and those infecting multiple hosts, it may represent an upper limit on divergence, beyond which homologous recombination in the six most conserved genes between two picornaviruses might not give viable progeny.

Introduction

Research in virology relies on virus taxonomy for providing a unified intellectual and practical framework for analysis, generalization, and knowledge dissemination. Despite its broad relevance, taxonomy has received relatively little attention from the research community. Virus taxonomy is developed under the direction of the Committee on Taxonomy of Viruses (ICTV) and recognizes five hierarchically arranged ranks: order, family, subfamily, genus, and species (in ascending order of intervirus similarity), with order and subfamily levels being used less commonly. Virus species are of principal importance³⁷³, and for their demarcation the so-called polythetic species concept^{29,465} is applied. Accordingly, viruses are recognized as single species if they share a broad range of characteristics while constituting a replicating lineage that occupies a particular ecological niche^{259,466}. These characteristics, so-called demarcation criteria, are devised for each genus separately and are revised periodically^{143,257}. To ensure that each virus is classified, they are allowed to vary greatly between and even within families, with no single unifying property being sought after (for a review, see reference⁴⁶⁷). Consequently, virus species are operational units that are delimited at the genus level. They can be contrasted to biological species that are commonly defined by shared gene pools and reproductive isolation. The lack of a mandatory common denominator of virus species casts uncertainty over the interpretation and generalization of results obtained across different genera.

We are interested in exploring the wealth of genomic information for improving the foundation of virus taxonomy. For this purpose, we used the family *Picornaviridae* as a case study. Picornaviruses form one of the largest and most actively studied virus families, with many human and societally important pathogens, whose number is steadily growing^{130,415}. They employ a single-stranded RNA genome of positive sense (ssRNA+) with lengths in the range of 6,500 to 9,000 nucleotides of which about 90% encode a single polyprotein that is co- and posttranslationally cleaved into 11 to 13 mature proteins³¹³. In total, six proteins, three of the capsid module (1B, 1C, and 1D, known also as VP2, VP3, and VP1), and three of the replicase module (2C, 3C, and 3D) are conserved family-wide to form the backbone of the genetic plan¹⁸². Other proteins may be specific for different subsets of picornaviruses. Particularly, proteins known as L and 2A come in a large variety of molecular forms^{182,264} most of which were implicated in functions that secure virus propagation in the host⁷. The open reading frame that encodes the polyprotein³⁴⁶ is flanked by the two untranslated regions, 5'-UTR and 3'-UTR. The 5'-UTR includes a highly structured internal ribosomal entry site (IRES) which is known to exist in five different molecular forms, from type I to type IV^{441,481}. The expert-based classification (the ICTV taxonomy) of the *Picornaviridae* devised by the Picornavirus Study Group (PSG), recognizes 28 species distributed among 12 genera and no subfamilies²⁶⁴. A growing number of picornaviruses either is tentatively classified in provisional taxa or remains unclassified. The PSG uses a complex set of rules to devise taxa and classify viruses. All genera form compact monophylogenetic clusters in separate

trees of the conserved proteins as well as the capsid and replicative modules, respectively. The polyprotein sequences of viruses in different genera differ by at least 58% amino acid (aa) residue identity^{263,435}. For genera that include multiple species (*Enterovirus*, *Cardiovirus*, *Aphthovirus*, *Parechovirus*, *Kobuvirus*, *Sapelovirus*), demarcation criteria that separate the species have been developed by the PSG. Most commonly, they define lower limits of pairwise amino acid identity in the polyprotein and its two parts, the capsid and replicative modules. Additionally, the criteria may include restrictions on genome organization, genome base composition (G+C), host range, host cell receptor variety, and compatibility in processes that underlie the replicative cycle. Some taxa may be distinguished by the presence of a molecular marker that could be an L and/or a 2A protein^{182,239}, the type of IRES^{205,481}, the genome position of internal cis-replicative element (CRE) directing the VPg synthesis^{81,438}, or a combination thereof. For genera that include a single species (*Hepatovirus*, *Erbovirus*, *Teschovirus*, *Senecavirus*, *Tremovirus*, *Avihepatovirus*), no species demarcation criteria have been developed due to the lack of sufficient diversity in the available virus sampling.

In an accompanying paper²⁸², we have introduced a quantitative approach for partitioning the genetic diversity of a virus family to build a hierarchical classification, which we named DEmARC (“DivErsity pARtitioning by hieRarchical Clustering”). In contrast to the framework of virus taxonomy, DEmARC uses a sole demarcation criterion—intervirus genetic divergence. When applying DEmARC to the family *Picornaviridae*, it clustered 1,234 genome sequences in groups at three hierarchical levels (the GENETIC classification). In this study, two of the three inferred levels in the GENETIC classification were found to correspond most closely to the species and genus ranks recognized by ICTV²⁶⁴. A few deviations from the ICTV taxonomy concern assignments for the genus *Aphthovirus*^{264,295} and species *Human rhinovirus A* and *C*^{19,427}. The third level has no counterpart in the current taxonomy. Furthermore, we found the family-wide conserved proteins to have almost universally accumulated fewer substitutions in viruses of the same species than in those belonging to different species, suggesting that picornavirus species are genetically separated. This also indicates that objective discrimination between the genetic divergence within a taxon (intragroup) and that between taxa (intergroup) is attainable. Finally, we outline conceptual differences between the frameworks that underlie the two classifications. These differences concern the relation of genetic diversity, the content of a genetics-based classification, and virus groups representing its structure. To facilitate the comparison, we introduce two novel diagrams that (i) illustrate the connection of the new approach developed in this study to conventional phylogenetic analysis already used in taxonomy and (ii) depict the classification and the associated genetic diversity in a standardized manner.

Materials and Methods

Virus sequences, multiple alignment, and distance estimation. Complete genome sequences for 1,234 picornaviruses available on 15 April 2010 at the National Center for Biotechnology Information GenBank/RefSeq³⁶ databases were downloaded using HAYGENS⁴²³ into the Viralis platform¹⁸³. A concatenated multiple-amino-acid alignment covering the family-wide conserved capsid proteins 1B, 1C, 1D and the nonstructural proteins 2C, 3C, and 3D of the 1,234 picornaviruses (Fig. 1) was produced using the MUSCLE program version 3.52¹²⁶, and poorly conserved columns were further manually refined. The alignment subsequently facilitated the calculation of pairwise evolutionary distances (PEDs) using a maximum likelihood (ML) approach^{66,145}, as implemented in the Tree-Puzzle program version 5.2⁴¹². The WAG amino acid substitution matrix⁴⁷⁸ was applied. PEDs serve as a measure of intervirus genetic divergence.

Phylogeny reconstruction. Bayesian posterior probability trees were compiled utilizing the Beast software version 1.4.7¹¹⁹. Bayesian Markov chain Monte Carlo (MCMC) chains (2 independent runs per data set) were run for 4 million steps (10% burning, sampled every 100 generations) under the WAG amino acid substitution matrix⁴⁷⁸. The substitution rate heterogeneity among alignment sites was allowed as modeled via a gamma distribution with 4 categories. The uncorrelated relaxed molecular clock approach (log-normal distribution)¹¹⁸ was used, as it was strongly favored over the strict molecular clock (log Bayes factor of 56.7) and the relaxed molecular clock approach with exponential distribution (log Bayes factor of 14.6). The convergence of runs was verified using Tracer version 1.4¹²⁰. ML trees were compiled utilizing the PhyML software version 3.0¹⁹⁶. The WAG amino acid substitution matrix was applied, and substitution rate heterogeneity among sites (4 categories) was allowed. Support values for internal nodes were obtained using the nonparametric bootstrap method with 1,000 replicates or through Shimodaira-Hasegawa (SH)-like approximate likelihood ratio tests.

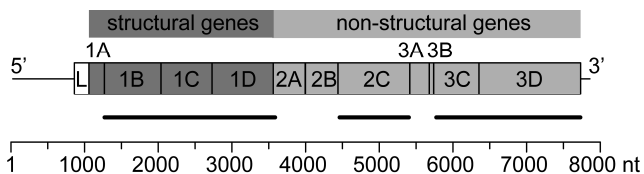


Figure 1. Picornavirus genome organization. The organization of the picornavirus genome is shown on the example of Porcine sapelovirus. Products derived after cleavage of the encoded polyprotein are indicated by rectangles and names. They include structural proteins (dark gray background) forming virus particles, nonstructural/accessory proteins (light gray) involved in replication and expression, and the leader protein (white), which is not found in all picornaviruses. The horizontal bars below highlight the six proteins conserved across the family. A concatenated, picornavirus-wide multiple alignment of these six proteins forms the data set of this study.

Genetics-based virus classification. We have developed DEmARC, a quantitative procedure for hierarchical classification of a virus family based on intervirus genetic divergence²⁸². It has been evaluated extensively for consistency and stability with respect to key parameters including the amount and/or diversity of the input data, the alignment construction method, and the measure of intervirus divergence. For brevity, we refer to the DEmARC-mediated picornavirus classification as the GENETIC classification.

Measures of quality. In the accompanying paper²⁸², we have introduced a cost measure to determine a threshold on intragroup genetic divergence at each classification level in a quantitative way. This cost is calculated as the cumulative violation of intragroup PED values to the respective threshold among all taxa of the level (see reference²⁸² for details). Hence, this cost, which is a nonnegative real number, is used as a quality measure for a classification level—the lower the cost the higher the quality. Furthermore, analogs of the cost measure can be calculated for both a taxon and a single virus by summarizing over the respective violating PED values.

Another measure of the quality of a taxon is the fraction of intraspecific pairwise distances not exceeding the distance threshold of the respective level, to which we refer as cluster quality (cq). A taxon is considered complete if the cq value is 1 and incomplete otherwise ($0 < cq < 1$).

Results and Discussion

Phylogeny, PED distribution, and classification of picornaviruses. Our data set included 1,234 genome sequences from picornaviruses whose taxonomic position at the start of this study was either already established as described above or remained provisional or uncertain due to the considerable time involved in taxa assignments²⁶⁴. Using a concatenated multiple alignment of six conserved proteins of a representative set of 38 picornaviruses, we reconstructed a phylogenetic tree under both an ML and a Bayesian framework. The two trees had a matching topology and included monophyletic branches corresponding to the taxa recognized by ICTV (Fig. 2, black tree branches and names). The phylogeny additionally comprised a number of new branches of different lengths accommodating a large number of relatively recently identified picornaviruses. We concluded that the alignment used in our study contains information compatible with taxonomy. Hence, we used this alignment as input for DEmARC in order to devise the GENETIC classification of picornaviruses²⁸². We identified three statistically most strongly supported positions of discontinuity (thresholds) in the picornavirus PED distribution that we assigned as defining species, genus, and supergenus levels of the classification.

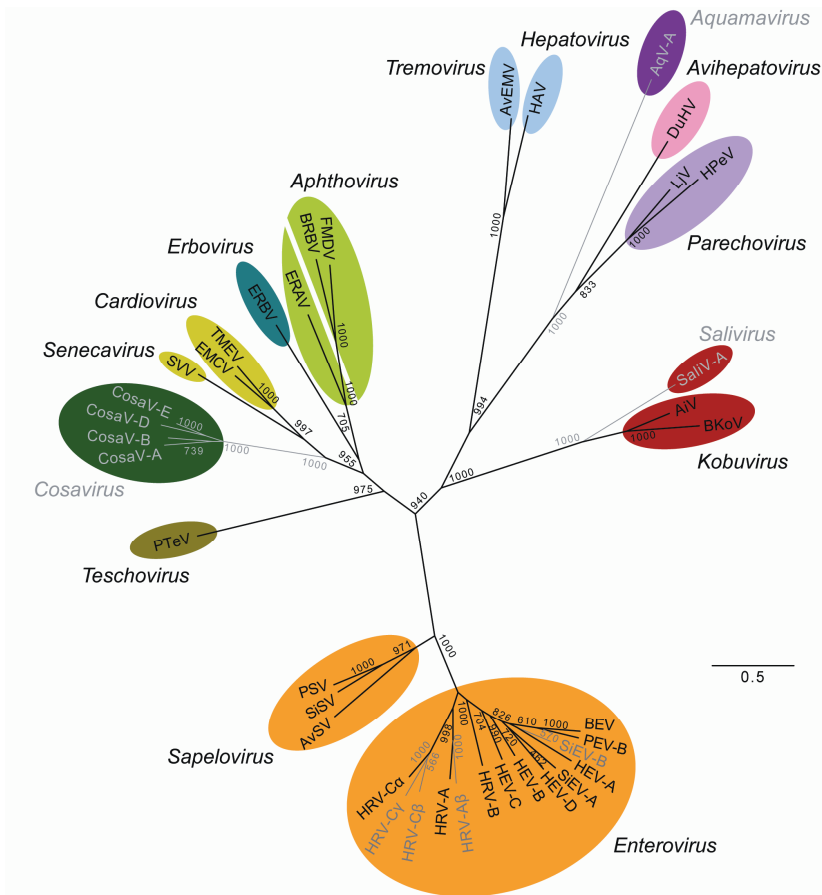


Figure 2. Phylogeny and GENETIC classification of the *Picornaviridae*. Shown is a maximum likelihood phylogeny of 38 picornaviruses representing species diversity based on the family-wide conserved proteins 1B, 1C, 1D, 2C, 3C, and 3D. A Bayesian analysis resulted in an identical tree topology (data not shown). The part of the tree representing the ICTV-defined 28 species and 12 genera is drawn in black, and provisional or currently not recognized taxa are in gray. Clusters equivalent to ICTV genera are highlighted by colored ovals. A split of *Aphthovirus* according to the GENETIC classification is indicated (white line). Genera with identical coloring unite to in total 11 supergenera identified in this study. The viruses shown represent the following species (*italics*) or species-like clusters according to the GENETIC classification: *Porcine sapelovirus* (PSV), *Simian sapelovirus* (SiSV), *Avian sapelovirus* (AvSV), *Human rhinovirus A* (HRV-A), *human rhinovirus Aβ* (HRV-Aβ), *Human rhinovirus B* (HRV-B), *human rhinovirus Cα* (HRV-Cα), *human rhinovirus Cβ* (HRV-Cβ), *human rhinovirus Cγ* (HRV-Cγ), *Human enterovirus A* (HEV-A), *Human enterovirus B* (HEV-B), *Human enterovirus C* (HEV-C), *Human enterovirus D* (HEV-D), *Simian enterovirus A* (SiEV-A), *Simian enterovirus B* (SiEV-B), *Porcine enterovirus B* (PEV-B), *Bovine enterovirus* (BEV), *Bovine kobuvirus* (BKoV), *Aichi virus* (AiV), *Salivirus A* (SaliV-A), *Human parechovirus* (HPeV), *Ljungan virus* (LjV), *Duck hepatitis A virus* (DuHV), *Aquamavirus A* (AqV-A), *Hepatitis A virus* (HAV), *Avian encephalomyelitis virus* (AvEMV), *Foot-and-mouth disease virus* (FMDV), *Bovine rhinitis B virus* (BRBV), *Equine rhinitis A virus* (ERAV), *Equine rhinitis B virus* (ERBV), *Theilovirus* (TMEV), *Encephalomyocarditis virus* (EMCV), *Seneca Valley virus* (SVV), *human cosavirus A* (CosaV-A), *human cosavirus B* (CosaV-B), *human cosavirus D* (CosaV-D), *human cosavirus E* (CosaV-E), *Porcine teschovirus* (PTeV). Numbers at branch points provide support values from 1,000 nonparametric bootstraps. The scale bar represents 0.5 amino acid substitutions per site on average.

Below, we compare the GENETIC classification and the ICTV taxonomy at each of these levels separately. To facilitate the comparison, we devised a special plot (Fig. 3A, middle), which connects the phylogeny (Fig. 3A, left) and the PED distribution (Fig. 3A, bottom right) that are used in taxonomy and DEmARC, respectively. The plot (Fig. 3A, middle) presents a two-dimensional partitioning of the intervirus genetic diversity. It reveals an association of a taxon in the tree and a range in the PED distribution that belongs to one of the three levels of the GENETIC classification. Thus, the phylogeny and the PED distribution represent complementary projections of the intervirus genetic diversity that, when combined, reveal the most critical characteristics utilized in taxonomy. The availability of this plot empowers the reader with a tool to inspect the foundations and analyze the implications of the proposed classification.

GENETIC classification versus ICTV taxonomy: species level. At the species level, the principal level in taxonomy, the GENETIC classification includes 38 clusters. Twenty-seven of them correspond one-to-one to species of the ICTV taxonomy⁴³⁵, three clusters encompass a single species (*Human rhinovirus C*; HRV-C), and eight clusters comprise recently discovered viruses that were not yet formally classified at the start of the study. HRV-C was split in three species-like clusters provisionally named Human rhinovirus C α (HRV-C α), Human rhinovirus C β (HRV-C β), and Human rhinovirus C γ (HRV-C γ) (Fig. 2 and 3A; Table 1).

The 27 clusters corresponding to the recognized species include already classified viruses and some accommodate also recently discovered viruses, including simian enteroviruses joining *Human enterovirus A* and *B* (HEV-A and HEV-B, respectively)^{340,342}, Saffold virus grouping with *Theilovirus*^{45,77,241,299}, possum enterovirus joining *Bovine enterovirus*⁴⁹⁷, and porcine kobuvirus being classified with *Bovine kobuvirus*³⁸⁶ (Table 1). With the exception of *Theilovirus*, the host range of these species was expanded as a result of this virus update. A recent phylogenetic study of RNA viruses from three families and two genera other than the *Picornaviridae* revealed that host switching by virus species is more frequent than previously thought²⁶².

The eight clusters encompassing exclusively novel viruses include the following: cosaviruses (4 clusters; CovaV-A, CovaV-B, CovaV-C, CovaV-D)^{221,249}, seal picornavirus (1 cluster; AqV-A)^{250,264}, human klasse- and saliviruses (hereafter referred to as saliviruses) (1 cluster; SaliV-A)^{190,222}, rhinoviruses close to but separated from *Human rhinovirus A*, HRV-A (1 cluster; provisionally named Human rhinovirus A β , HRV-A β)^{81,349,351}, and simian enteroviruses not belonging to *Simian enterovirus A* (1 cluster; SiEV-B)^{338,340,342} (Table 1). There seems to be a good match between the GENETIC classification assignments listed above and those that are in the pipeline for approval by ICTV^{263,264}.

Thirty-two out of 38 species include more than one sequence (nonsingleton). Few of these determine the PED range of all 38 species clusters, which is defined as “intraspecies” genetic divergence (Fig. 3A).

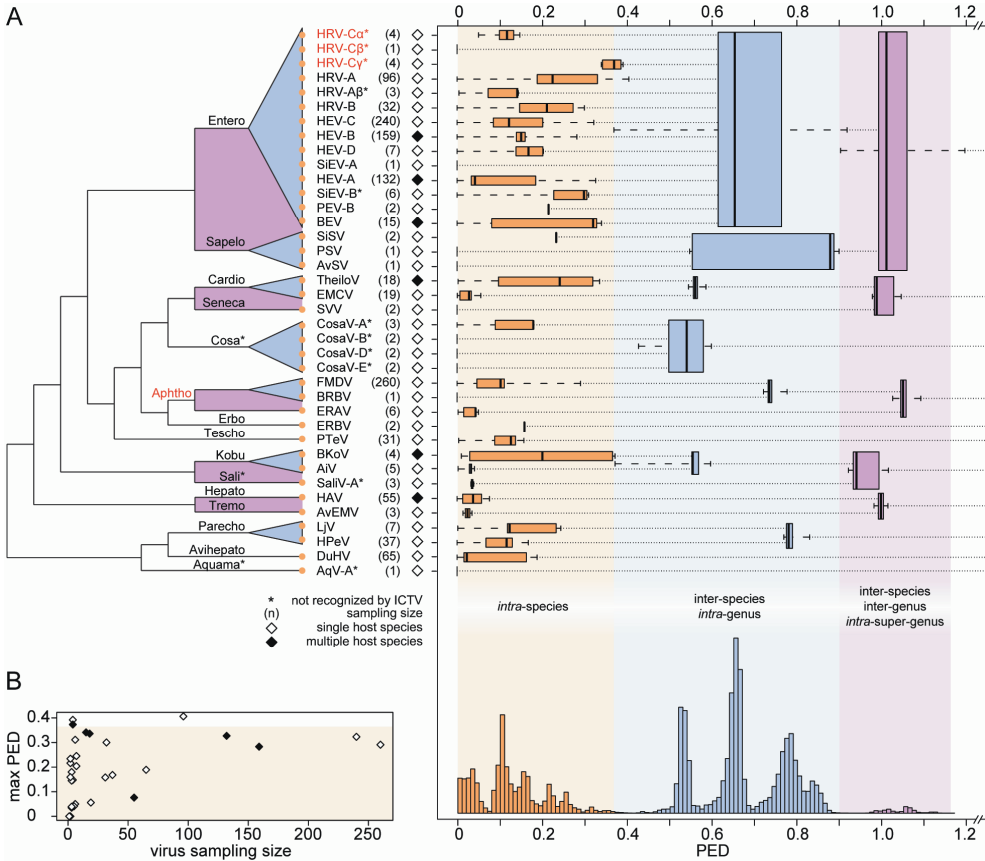


Figure 3. Intragroup genetic divergence and species sampling size. (A) Box-and-whisker graphs were used to plot distributions of distances between viruses from the same species (orange), between viruses from different species but the same genus (blue), and between viruses from different genera but the same supergenus (purple). The boxes span from the first to the third quartile and include the median (bold line), and the whiskers (dashed lines) extend to the extreme values. For name abbreviations, see the Fig. 2 legend; numbers in brackets correspond to the number of sequences per species; open and filled diamonds indicate single and multiple host species range, respectively. Genera and supergenera constituting only one species are not shown. The corresponding first half of the PED distribution (see reference²⁸²) is depicted below. Phylogenetic relationships of the 38 picornavirus species are shown by the cladogram to the left (following the topology in Fig. 2) with intragenus relations collapsed. Colored shapes indicate those taxa that contribute to intragroup distances to the right. Species and genera currently not recognized by ICTV are marked with asterisks, and discrepancies between the ICTV taxonomy and the GENETIC classification (not caused by recently discovered viruses) are highlighted in red. (B) The relationship between sampling size and maximum intragroup genetic divergence is shown for each species.

Table 1. Differences between GENETIC classification and ICTV taxonomy on the species level.

Virus ^a	Difference			
	Type ^b	ICTV ^c	GENETIC ^d	Quality ^e
Simian picornavirus 17	new	-	HEV-B	1
Simian picornavirus 13	new	-	HEV-A	1
Simian enterovirus SV19, SV43	new	-	HEV-A	1
Saffold virus	new	-	TheiloV	1
Possum enterovirus W1, W6	new	-	BEV	1
Seal picornavirus type 1	new	-	AqV-A*	-
Simian enterovirus N125, N203, SV6	new	-	SiEV-B*	1
Enterovirus 103 isolate POo-1	new	-	SiEV-B*	1
Human cosavirus A1, A2	new	-	CosV-A*	1
Human cosavirus B	new	-	CosV-B*	-
Human cosavirus D	new	-	CosV-C*	-
Human cosavirus E	new	-	CosV-D*	-
Salivirus NG-J1, Human klassevirus 1	new	-	Saliv-A*	1
Porcine kobuvirus S-1-HUN, K-30-HUN	new	-	BKoV	0.833
Human rhinovirus VR-1118, VR-1155, VR-1301	new	-	HRV-A β *	1
Human rhinovirus C 026, NY-074, NAT001, QPM	mm	HRV-C	HRV-C α *	1
Human rhinovirus C 025	mm	HRV-C	HRV-C β *	-
Human rhinovirus C N4, N10, NAT045	mm	HRV-C	HRV-C γ *	0.500

^a Shown is the Definition field value in the Genbank annotation of one or several viruses.

^b A virus was not available or assigned to a tentative species at time of the ICTV release (new); a mismatch was observed between the ICTV taxonomy and GENETIC classification (mm).

^c It is shown to which species the virus is classified in the ICTV taxonomy; -, not available at the time.

^d It is shown to which species the virus was assigned in the GENETIC classification; new species proposed by the GENETIC-classification are indicated using asterisks. For species abbreviations, see Fig. 2 legend.

^e The proportion of intraspecies PED values not exceeding the species distance threshold; -, for clusters with less than 3 viruses.

Virus sampling for the 38 species varied considerably in a range of 1 (six species) to 260 (*Foot-and-mouth disease virus* [FMDV]) sequences (Fig. 3). The corresponding intragroup PED ranges (distances between virus pairs belonging to a single species) differed ~10-fold among the species with more than one nonidentical sequences, with maxima varying from 0.04 (avian encephalomyelitis virus [AvEMV]) to 0.41 (HRV-A) (Fig. 3A). All except three species clusters were complete (each intragroup PED is below the species distance threshold) (Fig. 3A) (see reference ²⁸²). The three incomplete species clusters included viruses that belong to HRV-A (96 viruses in total and 14 viruses define pairs with larger-than-threshold distances), *Bovine kobuvirus* (4 and 1), and the proposed species-like cluster HRV-C γ (4 and 2) (Table 2; Fig. 4). In these species, respectively, 3.6%, 16.7% and 50% of intragroup PEDs exceeded the species threshold (Table 1; Fig. 3A). Combined, they account for less than 0.19% (175 out of 93,857) of all intragroup PED values at this level. One of these, *Bovine kobuvirus* was split in two clusters that observe the threshold and are host restricted in our analysis of three evaluation data sets²⁸². This splitting would be in line with the original proposal by the authors who identified the porcine kobuvirus³⁸⁶.

Table 2. Violations to a distance threshold in the GENETIC classification.

Accession no.	Virus ^a	Threshold	Violations ^b	Cost ^c
FJ445152	Human rhinovirus 71, ATCC VR-1181	Species	33	0.902
FJ445136	Human rhinovirus 51, ATCC VR-1161	Species	17	0.770
GQ415052	Human rhinovirus A, hrv-A101-v1	Species	16	0.707
FJ445147	Human rhinovirus 65, ATCC VR-1175	Species	14	0.577
FJ445156	Human rhinovirus 80, ATCC VR-1190	species	14	0.431
GQ415051	Human rhinovirus A, hrv-A101	Species	13	0.434
FJ445120	Human rhinovirus 20, ATCC VR-1130	Species	13	0.393
DQ473507	Human rhinovirus 53	Species	11	0.285
FJ445150	Human rhinovirus 68, ATCC VR-1178	Species	11	0.187
DQ473508	Human rhinovirus 28	Species	10	0.255
DQ473506	Human rhinovirus 46	Species	6	0.154
FJ445183	Human rhinovirus 78, ATCC VR-1188	Species	6	0.149
EF173418	Human rhinovirus 78	Species	6	0.130
DQ473497	Human rhinovirus 23	Species	1	0.003
NC_009996	Human rhinovirus C	Species	2	0.100
EF077280	Human rhinovirus NAT045	Species	1	0.049
NC_004421	Bovine kobuvirus	Species	1	0.011
AF119795	Enterovirus 71, TW/2272/98	Genus	21	0.157
NC_006553	Avian sapelovirus	Supergenous	7	0.195

^a Definition field value in the Genbank annotation; viruses of the same taxon are separated from others by an empty row. Only the minimal subset of violating viruses sufficient to explain all violating PEDs are listed.

^b Number of PEDs exceeding the respective distance threshold.

^c Cumulative value of the disagreement of a virus to the respective distance threshold; calculated as the virus-specific clustering cost (see reference ²⁸²) using the threshold as a unit.

GENETIC classification versus ICTV taxonomy: rhinoviruses. Why do the GENETIC classification and the ICTV taxonomy differ so profoundly in respect to HRV-C while agreeing on the virus composition of all other species? Specifics of both HRV-C evolution and the two classification frameworks could play a role. The genetic diversity of these viruses in capsid (1A, also known as VP4, and 1D proteins) and nonstructural (3D) regions was previously reported to exceed those of other rhinoviruses^{318,427}. In the 1D protein, this difference is smallest, and the entire HRV-C diversity was considered to be below the species divergence limit, paving the way for the recognition of HRV-C as a single species. We have also observed HRV-C viruses to form a single species-like cluster in the DEmARC-mediated classification using the major capsid proteins only²⁸². However, in the analysis of the data set comprising the six family-wide conserved proteins, the observed maximum divergence of HRV-C considerably exceeded that of its most diverged subset (HRV-Cγ) and the family-wide species demarcation threshold: 0.424, 0.392, and 0.37, respectively. This was likely due to an accumulated effect of compatible phylogenetic signals from both the structural and the nonstructural proteins (Fig. 4 and data not shown). The virus divergence in HRV-C is so high that even half of intragroup distances in HRV-Cγ exceed the species threshold (Fig. 3A; Table 1). This low support for the HRV-Cγ species (Table 1), which is the lowest overall and only one of three below 100%, is even more striking given that the virus

sampling in this provisional species and the two HRV-C sister taxa is very limited (one to four available genome sequences per cluster). Thus, it remains plausible that with the accumulation of sequenced genomes in the future, HRV-Cγ will be split further, increasing the number of provisional HRV-C species to at least four compared to the one currently recognized. Each of these species corresponds to a separate major lineage in the HRV-C phylogeny³¹⁸ (Fig. 4).

Furthermore, the GENETIC classification proposes the recognition of another potentially new rhinovirus species (HRV-Aβ). It is formed by three viruses and corresponds to the recently identified “clade D” rhinoviruses³⁵¹ (known otherwise as the cluster HRV-A2⁸¹) that is a sister group to the species HRV-A (Fig. 4). Altogether, our analysis suggests that at least six (rather than three) human rhinovirus species may exist. Testing this more complex species structure in human rhinoviruses could facilitate research into the molecular basis of the observed clinical heterogeneity of rhinovirus infections in humans^{19,234,349}.

GENETIC classification and recognition of virus species as biological entities. We have found that viruses belonging to a single species are usually separated by less than ~0.4 replacements per residue on average in the six most conserved proteins, while this distance is commonly exceeded in virus pairs representing different species (Fig. 3B). Furthermore, we observed a dependence of the largest intragroup genetic divergence (maximum intragroup PED) on the sampling size (number of viruses) in the 38 species: with increasing sampling size, a species' maximum genetic divergence tends to approach the species distance threshold (Fig. 3B). Accordingly, the 11 species that constitute the upper ~25% of the maximum PED range are enriched with highly sampled species. Additionally, host range may be another parameter of relevance to the genetic divergence of species: the upper ~25% of the maximum PED range is also enriched with species that infect multiple hosts (five out of six species of this kind) (Fig. 3B). This correlation is sensible biologically, since host switching is expected to be accompanied with accelerated virus evolution.

The above-mentioned correlations involve species that belong to four genera, indicating that they may be applicable to all picornavirus species. If so, we may expect that with a sufficient increase of the species sampling size, the maximum divergence of all species in the *Picornaviridae* will approach the species threshold. This would indicate that the intragroup genetic divergence of species is constrained similarly in different lineages. Alternatively, some currently undersampled lineages could accommodate a smaller natural diversity due to either stricter constraints or being a “young” species. For instance, *Hepatitis A virus* with its relatively large sampling size and two hosts (Fig. 3B) has an unusually small maximum genetic divergence (see also reference³¹). Thus, it remains possible that the inferred species threshold represents an upper limit on the maximum intragroup genetic divergence but that the actual limit may be smaller in some picornavirus species. Likewise, we may not exclude that viruses in some species may diverge above the threshold. This might happen due to position-specific variations of replacements in the six conserved

proteins or involvement of virus lineages that are in the transition to the establishment of separate species. The virus diversity known in taxonomy as the species *Human rhinovirus A* and *Human rhinovirus C* (Fig. 4) could represent such cases. Also, it is important to stress that the species distance threshold represents an average of over 2,446 positions in six conserved proteins²⁸² indicating that (lineage-specific) variations of maximum divergence for different proteins are likely (see below and also references^{293,425}). Further characterization of the natural diversity of picornavirus species, including the surveillance of novel hosts, could address this important aspect of the species delimitation in the GENETIC classification.

The existence of a species threshold on intragroup genetic divergence must be rationalized mechanistically. It may be a manifestation of speciation due to changes accumulated in either conserved proteins or other elements encoded in the picornavirus genome. To discuss the alternatives, it is important to recall that the divergence is a net result of contributions from several sources, including mutation and homologous recombination. Although both promote diversity increase, they act in opposite directions concerning progeny divergence: on average, the progeny of two lineages diverged by mutation will be more separated than their parents, while those generated through homologous recombination of parents will be closer to each other than to their parents³⁰⁵. In other words, recombination limits the maximum genetic divergence in an asexual population; without it, the population will evolve into separate, more distantly related lineages after a sufficient time.

The inferred species threshold reflects the maximum amount of accumulated genetic differences in the six conserved proteins between two picornaviruses that remains compatible with the viability of progeny produced by homologous recombination, as argued below. The frequency of homologous recombination depends on the extent of base pairing, with intratypic recombination being most common^{260,456}. Two picornaviruses that are separated by a distance approaching the species threshold would retain only relatively small stretches of identical orthologous residues in their genome because the threshold is so high; the lack of extensive base pairing should impede homologous recombination. Even if recombination happens between these viruses, the resulting chimeric progeny will be viable only if the recombinant proteins, which all are essential for virus reproduction, remain functional. The protein functionality depends on the intra- and interprotein compatibility of lineage-specific mutations that have been accumulated since the divergence of these viruses. The mutation spectrum is restricted by so-called epistatic interactions between different protein positions⁴²⁰, making mutations outside this spectrum incompatible with the protein functioning. As two viruses diverge, they will approach the species distance threshold beyond which accumulated mutations may become incompatible with progeny viability in any combination that could be generated in the recombinants. In this framework, the existence of the species threshold reflects the genetic separation of species. This model could be probed in experiments on virus chimeras involving the conserved backbone

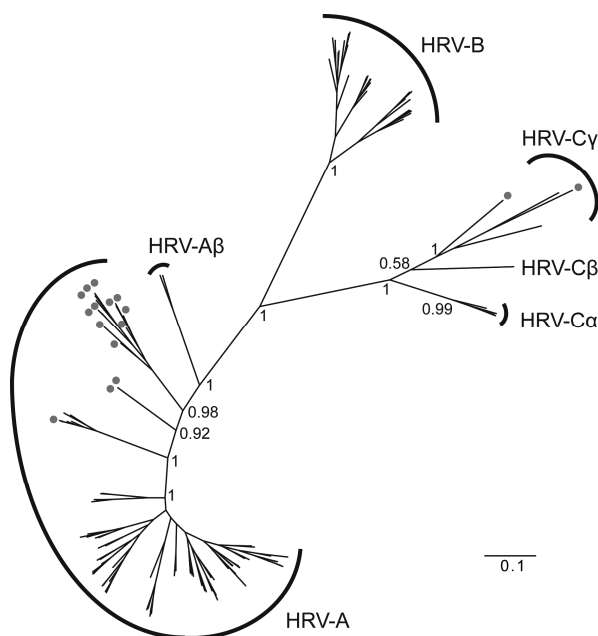


Figure 4. Phylogeny of rhinoviruses. Shown is an ML phylogeny for 140 rhinoviruses based on the family-wide conserved proteins 1B, 1C, 1D, 2C, 3C, and 3D. SH-like support values are shown for basal branching events. Species taxa recognized by the GENETIC classification are indicated (see also the Fig. 2 legend). A minimal set of viruses sufficient to explain all violating PEDs that exceed the species distance threshold are highlighted by gray dots (see Table 2 for details on involved viruses). The scale bar represents 0.1 amino acid substitutions per site on average.

proteins. It is predicted that intra- but not interspecies chimeras must be viable. Results compatible with this model are available for *Human enterovirus C*^{235,238}. The viability of chimeric progeny may be determined not only by the distance between parents but also by the origins of combined parts²³⁸, indicating that both reciprocal chimeras must be characterized.

In the alternative model, other elements outside of the conserved proteins could be implicated in the control of speciation. These elements include L and 2A proteins, which exist in a large variety of molecular forms in picornaviruses^{7,182,264}, or CRE, whose location in the genome varies tremendously among picornaviruses^{81,156,163,182,313,464,485,486}, or other elements located in the 5' and 3' noncoding regions^{438,481}. For a number of picornaviruses, the viability of interspecies chimera carrying a noncognate version of either L³⁶⁴ and 2A protein³⁰³ or CRE⁴⁶⁴ and IRES^{304,481} was demonstrated experimentally. Also, several picornaviruses with deleted L proteins were found to be viable^{266,365}, which is in line with their accessory “security” role in virus replication⁷. Thus, picornaviruses could accept “gene flow” from other species in the case of elements that are not conserved family-wide. Consequently, an acquisition or loss or relocation of a nonconserved element by a

picornavirus in vivo seems plausible. Furthermore, it is conceivable that such a newly acquired element might confer a function that would allow the virus to explore a new niche, eventually leading to its reproductive isolation from other lineages; in other words, it would trigger speciation. However, this model does not provide a mechanistic explanation for the species genetic threshold other than that of the first model (see above).

Thus, in our opinion, nonconserved and conserved elements of the picornavirus genome may play distinct roles in speciation. The clear-cut relation between the species delimitation and the discontinuity in the intervirus genetic distance distribution lends support to the notion that picornavirus species are biological entities rather than merely operational units.

GENETIC classification versus ICTV taxonomy: genus level. The GENETIC classification includes a genus level comprising 16 clusters. Eleven of them match ICTV genera, two clusters encompass a single genus (*Aphthovirus*), and three clusters comprise recently discovered viruses (Fig. 2 and 3A). The genus *Aphthovirus* was split into two clusters that are formed by the single species *Equine rhinitis A virus* (ERAV)²⁹⁵ and the two species *Foot-and-mouth disease virus*^{109,265} and *Bovine rhinitis B virus* (BRBV)²¹⁵, respectively. The minimum PED of 1.03 between viruses of these two clusters is considerably larger than the genus distance threshold of 0.905 and comparable to those between the closest virus pairs of other sister genera, e.g., *Senecavirus* and *Cardiovirus* or *Enterovirus* and *Sapelovirus*. In fact, the distance range between viruses of these two clusters fits in the limits of the next rank (supergenous) that is considered below. This result was also reproduced in classifications of two evaluation data sets²⁸² in which these viruses are present but which differed in respect to genome region and virus selection, respectively. We note that an L protein variety with a papain-like fold and proteolytic activity that is associated with this monophyletic virus group²⁶⁴ could be considered a molecular marker of a larger group that also includes the sister genus *Erbovirus*^{295,484}. Thus, there is a strong support for splitting the genus *Aphthovirus* into two genera in future revisions of taxonomy.

The three genus clusters that are formed by recently discovered viruses include cosaviruses (4 species), seal picornavirus (1 species), and saliviruses (1 species). All genus clusters were complete with the exception of *Enterovirus* (Fig. 3A) resulting in less than 0.02% (21 out of 152,194) of intragroup PED values that exceed the genus threshold (Table 2), all involving a single sequence of enterovirus 71 (GenBank accession number, AF119795) from HEV-A. Seven out of 16 genera are nonsingletons. Few of these determine the genus-specific PED range, which is defined as “interspecies intragenus” genetic divergence (Fig. 3A).

GENETIC classification versus ICTV taxonomy: recognition of the new hierarchical level supergenous. The GENETIC classification recognizes an additional rank—provisionally called supergenous—that has no counterpart in virus taxonomy. The threshold support for

this level is the strongest overall²⁸², indicating that it may reflect a clustering that is genetically and evolutionary sensible. At this level, we observed five nonsingleton supergenera that included more than one genus. They included viruses from 28 species and 10 genera. Four of these supergenera represented unions of, respectively, *Enterovirus* with *Sapelovirus*, *Cardiovirus* with *Senecavirus*, *Hepatovirus* with *Tremovirus*, and *Kobuvirus* with the cluster formed by recently discovered saliviruses (Fig. 2 and 3A). The fifth nonsingleton supergenus corresponds to the genus *Aphthovirus* in the ICTV taxonomy, which is split in two genera in the GENETIC classification (see above). The other six supergenera accommodate singleton genera, including 10 species in total. Four of these supergenera, *Avihepatovirus*, *Erbovirus*, *Parechovirus*, and *Teschovirus*, include only a single ICTV genus. Two supergenera are formed by recently discovered cosaviruses and seal picornavirus, respectively. All supergenus clusters are complete with the exception of the *Enterovirus/Sapelovirus* union (Fig. 3A), resulting in less than 0.25% (7 out of 2,814) of intragroup PED values that exceed the supergenus threshold (Table 2), all involving a single sequence of avian sapelovirus (RefSeq accession NC_006553) from AvSV. The five nonsingleton supergenera determine the supergenus-specific PED range, which is defined as “interspecies intergenus intrasupergen” genetic divergence (Fig. 3A).

Multimodality of PED distribution and evolution of picornaviruses. To our knowledge, there is nothing in evolutionary theory that would predict the multimodality of the PED distribution of conserved proteins for a virus family. However, once observed, it requires an (evolutionary) explanation. The model of virus speciation outlined above may explain the existence of PED discontinuity in which the species threshold resides. This threshold is expected to limit intragroup but not intergroup genetic divergence of lineages once they have crossed the threshold. This biological reasoning seems not to be applicable to other areas of PED discontinuity that are associated with the genus and supergenus thresholds. One plausible explanation for these discontinuities is that they could reflect large-scale changes in the rates of birth and death that might have happened across all virus lineages. Cellular life forms are known to have gone through alternating periods of both mass birth and death across lineages^{382,416}. If ancestral (picorna)viruses followed their hosts, alternating peaks and valleys in their PED distribution would reflect periods characterized predominantly by virus speciation and extinction, respectively. Thus, the genus and supergenus levels determined in this study would correspond to two major waves of speciation that are separated by two waves of extinction in the evolution of picornaviruses, possibly reflecting changes in the environment.

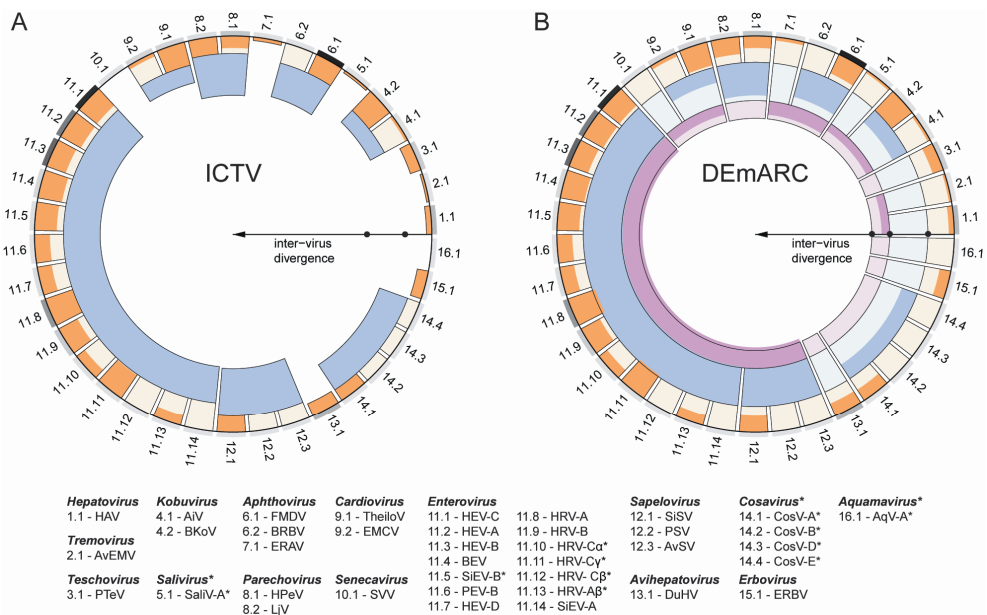
GENETIC classification and taxonomy of picornaviruses: two different perspectives on known and unknown virus diversities. As shown above, there is striking agreement between the GENETIC classification and the ICTV taxonomy⁴³⁵ of the *Picornaviridae* at the species and genus levels, with notable differences concerning the recognition of only few

taxa. The observed match is nontrivial⁴⁶⁷, since the underlying decision-making frameworks seek to satisfy different criteria. To fully reveal an impact of these criteria in the two frameworks, which are either exclusively (DEmARC) or predominantly (ICTV) genetics based, we sought to characterize their effect on partitioning the virus diversity, the primary target of classification and an important subject of research in virology. To this end, we have developed a circular diagram for presenting the classification of a virus family in a graphical form (Fig. 5). It depicts the proportions of the intervirus genetic divergence that is partitioned and not partitioned by a classification, respectively. The circle radius is defined by the PED range observed in the family, with intervirus genetic divergence increasing linearly from the perimeter (PED of zero) toward the center of the circle (maximum observed PED). Taxa are shown as boxes with heights (in radial dimension) that correspond to the PED range of the respective classification level. Species form the most external layer, followed by the genus layer, and—for the GENETIC classification—the supergenus layer residing closest toward the circle center. Within each taxon, the PED range that has been sampled and not sampled is colored according to the coloring scheme for classification ranks (Fig. 3) using bright and soft colors, respectively. The PED range that has not been partitioned (yet) by a classification (inner part of the circle) is in white.

To facilitate an unbiased comparison of the genetic foundations of both frameworks involving as many taxa as possible, the ICTV taxonomy in Fig. 5 was required to follow the GENETIC classification by accepting all taxa containing new viruses and those two (*Aphthovirus* and *Human rhinovirus C*) that were classified differently. As a result, the taxonomy and the GENETIC classification match each other in relation to the virus sampling per taxon (Fig. 5A and B, the most external layer) and the species and genus structure. At the species level, the PSG applies demarcation criteria that are genus specific and determined by the maximum observed intragroup genetic divergence among all sampled species of the genus. As a consequence, the limit on intragroup genetic divergence of species varies tremendously between genera. Accordingly, in the ICTV diagram only species of the same genus have equal heights (Fig. 5A, compare taxa 11.x with 12.x); for species that comprise a single virus, the height is nil (no pair is available to produce a PED; for instance, taxon 16.1 in Fig. 5A). At the genus level, the PSG does not provide demarcation criteria for the quantification of maximum intragroup genetic divergence and each genus is demarcated separately, usually by means of standard phylogenetic analyses. To reflect this approach, we represented genera as boxes whose heights correspond to the maximum observed intragroup genetic divergence (Fig. 5A). For genera comprising a single species the height, is nil (see for instance taxon 15.1 in Fig. 5A). In contrast, in the DEmARC diagram (Fig. 5B), all species, genus, or supergenus taxa have uniform, level-specific heights, since in this framework family-wide limits on intragroup genetic divergence are devised (compare for instance taxa 10.1 and 11.1 in Fig. 5B).

As a consequence of the utilization of family-wide demarcation thresholds, the DEmARC framework, compared to that of ICTV, partitions a larger share of the total PED

space (compare the sizes of white areas in Fig. 5A and B). Additionally, DEmARC unravels the intragroup genetic divergence ranges that might have been reached but remain to be described for most taxa (Fig. 5B, soft-colored areas). Such predictions are not available in the ICTV framework. The diagrams also reveal that most distant relations of viruses in the *Picornaviridae* remain totally unstructured (Fig. 5A and B, white central area). In the DEmARC framework, this area is smaller because it is partially partitioned by supergenera. It could be partitioned further if the subfamily level is introduced²⁸².



Concluding remarks. In a field lacking a gold standard, the striking agreement between the GENETIC classification and the expert-based taxonomy^{263,435} of the *Picornaviridae* could be seen as a cross-validation for both. Of principal importance is that the observed agreement implies that genomes may contain necessary and sufficient information to build a (picorna)virus taxonomy by using an approach²⁸² that employs a sole (rather than polythetic) demarcation criterion. There are additional benefits of the single criterion: its utilization provides consistency across all taxa, defines expected divergence ranges for poorly sampled taxa, reveals problematic taxa, and makes taxonomy fully genetics based. We expect the latter to facilitate the interaction between taxonomy and fundamental and applied research. Genetically delimited taxa could be readily targeted for recognition by virus diagnostics. Furthermore, the validity of the species threshold could be probed in experiments involving homologous recombinants in the backbone genes as well as through characterization of the natural virus diversity in already established and newly identified picornavirus species. Biological foundations of other, higher-rank thresholds could also be addressed. These advancements, combined with the application of DEmARC to other virus families, could bring virus taxonomy into the mainstream of research and pave the way to ultimately unite it with the taxonomy of cellular life forms.

Acknowledgments

We are indebted to Igor Sidorov, Andrey Leontovich, and Ivan Antonov for helpful discussions and suggestions and Dmitry Samborskiy, Igor Sidorov and Alexander Kravchenko for administrating and advancing different Viralis modules. This work was partially supported by the Netherlands Bioinformatics Centre (BioRange SP 2.3.3), the European Union (FP6 IP Vizier LSHG-CT-2004-511960 and FP7 IP Silver HEALTH-2010-260644), the Collaborative Agreement in Bioinformatics between Leiden University Medical Center and Moscow State University (MoBiLe), and Leiden University Fund (Special Chair in Applied Bioinformatics in Virology).

