Partitioning the Genetic Diversity of a Virus Family: Approach and Evaluation through a Case Study of Picornaviruses

Chris Lauber^a and Alexander E. Gorbalenya^{a,b}

Molecular Virology Laboratory, Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands,^a and Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia^b

The recent advent of genome sequences as the only source available to classify many newly discovered viruses challenges the development of virus taxonomy by expert virologists who traditionally rely on extensive virus characterization. In this proof-of-principle study, we address this issue by presenting a computational approach (DEmARC) to classify viruses of a family into groups at hierarchical levels using a sole criterion—intervirus genetic divergence. To quantify genetic divergence, we used pairwise evolutionary distances (PEDs) estimated by maximum likelihood inference on a multiple alignment of family-wide conserved proteins. PEDs were calculated for all virus pairs, and the resulting distribution was modeled via a mixture of probability density functions. The model enables the quantitative inference of regions of distance discontinuity in the family-wide PED distribution, which define the levels of hierarchy. For each level, a limit on genetic divergence, below which two viruses join the same group, was objectively selected among a set of candidates by minimizing violations of intragroup PEDs to the limit. In a case study, we applied the procedure to hundreds of genome sequences of picornaviruses and extensively evaluated it by modulating four key parameters. It was found that the genetics-based classification largely tolerates variations in virus sampling and multiple alignment construction but is affected by the choice of protein and the measure of genetic divergence. In an accompanying paper (C. Lauber and A. E. Gorbalenya, *J. Virol.* 86:3905–3915, 2012), we analyze the substantial insight gained with the genetics-based classification approach by comparing it with the expert-based picornavirus taxonomy.

Viruses form a large class of biological entities of extreme diversity (18). Unlike cellular organisms, they share neither a single common gene nor any other universally conserved trait that can be used to infer their phylogeny. This comes along with profound consequences and has resulted in a distributed approach to virus taxonomy adapted by the virological community. It is developed and advanced by independent study groups (SGs) on different viruses (see below) that operate under the auspices of the International Committee on Taxonomy of Viruses (ICTV) (26, 35). Virus taxonomy recognizes five hierarchically arranged ranks: order, family, subfamily, genus, and species (in ascending order of intervirus similarity). Only a relatively small subset of viruses is classified in subfamilies and/or orders, while the use of other ranks is most common.

The traditional development of virus taxonomy by SGs has been challenged by a growing gap between virus discovery and virus characterization. In this respect, genome sequences have been increasingly explored by practitioners. This line of research is driven by several developments. Essentially, all known viruses have their genomes sequenced largely due to the significant advances in sequencing techniques and the associated fall of costs over the last few years (7, 56). For a growing number of viruses, the genome sequence is the first and often the only information available (for a review, see references 13, 15, and 22). Successful incorporation of these viruses into the taxonomy framework through genome-based analyses has stimulated practice and research in extending this effort to all viruses, including those whose phenotypes have been probed. To recognize a taxon and/or classify a virus, it is common to seek a monophyletic group in a tree whose viruses could preferably be distinguished from other viruses by the possession of a unique molecular characteristic (marker) that thus can serve as a criterion for classification (35).

Another complementary approach that is steadily growing in popularity is so-called *pairwise sequence comparisons* (pasc) (60).

This approach utilizes a frequency distribution of pairwise sequence divergence between viruses to identify ranks and taxa (Fig. 1). Recognizing its broad utility in virology, a Web-based implementation of pasc, called appropriately PASC, was launched at the National Center for Biotechnology Information (NCBI) (6). Over the years, and mostly during the last decade, pasc has been used to propose, update, or revise the taxonomy of several virus families or genera (1–3, 8, 16, 25, 28, 44, 48, 58, 60, 71).

The current practice in pasc applications has three aspects in common. First, researchers typically seek to build a hierarchical classification with an *a priori*-defined number of levels that match usually the species and genus ranks of taxonomy. This approach normally guarantees a solution, but complexities of intervirus relations may remain not fully explored. Second, classification levels are delineated by imposing thresholds on the limits of intragroup genetic similarities at each level. How these thresholds are identified remains largely a matter of expert decision that places the thresholds outside a statistical framework and casts uncertainty about their validity. Third, observed identity percentages are commonly used for virus comparison. Their calculation is technically straightforward and fast. However, the applicability of this measure to data sets with considerable genetic divergence may be compromised by saturation effects that are linked to multiple substitutions at a site (41). Since RNA viruses are known for the extremely high mutation rates of their polymerases (19, 20, 67),

Received 20 December 2011 Accepted 13 January 2012

Published ahead of print 25 January 2012

Address correspondence to Alexander E. Gorbalenya, a.e.gorbalenya@lumc.nl.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JVI.07173-11

FIG 1 Grouping viruses based on thresholds in the distribution of pairwise genetic divergence. Shown is a fictitious example involving eight viruses that illustrates the relation between the selection of a threshold in the distribution of intervirus genetic divergence and the accompanied change in virus grouping. (A to D) An undirected graph representation is used to show viruses (black dots), virus groups (gray ovals), and pairwise genetic divergence between viruses of the same group (colored lines). Groups are defined as connected components of the graph which are formed by connecting those virus pairs (blue edges) whose divergence does not exceed a given threshold. Some intragroup divergence values may exceed the threshold (violations; purple edges). (E to H) The same data as on top, now shown as a frequency distribution (histogram) of genetic divergence between all virus pairs with four different divergence thresholds (dashed vertical line). Intragroup divergence values obeying a threshold are shown in blue, and those violating it are shown in purple. Intergroup divergence is in white. (A and E) A trivial clustering in which the number of virus groups equals the number of viruses. No pairwise divergence values are utilized. (D and H) The second trivial clustering in which all viruses join a single virus group. All pairwise divergence values are utilized. (B and F) A nontrivial clustering consisting of three virus groups for which eight intragroup divergence values obey the threshold and three violate it. (C and G) Another nontrivial clustering consisting of two virus groups for which only a single intragroup divergence value violates the threshold. Typically, the choice of a threshold is subjective in current practice. In this study, we show (see Materials and Methods) that the violating divergence values (F and G) can be used to define a cost for an applied divergence threshold, and we apply this measure to rank thresholds. Accordingly, thresholds resulting in a lower cost are favored, which makes the clustering in C superior to that in B. This simplified example illustrates how a classification at a single level is derived (the trivial solutions in A and D are not considered). As detailed in Materials and Methods, the approach outlined above can be separately applied to multiple divergence thresholds (each at a different location in the distribution), which would result in a hierarchical classification of the viruses.

pairwise identity percentages may indeed misrepresent the actual distances between the viruses. In addition to the above-mentioned common elements, pasc applications vary in respect to a number of parameters. The identity values may be calculated for either nucleotide or deduced amino acid sequences and be compiled on either pairwise or multiple sequence alignments. In some studies, only single genes/proteins were used, whereas others analyzed either multiple (concatenated) genes/proteins or complete genomes. How these specific choices and commonalities of the various pasc applications affect the end result remains a largely unexplored territory. This may be of relatively small concern as long as pasc results remain one of several characteristics in decision-making in virus taxonomy. However, with the current trend to follow the results of pasc-based analyses, its practice and quality may soon become dominant factors in taxonomy without having been evaluated properly.

In this study, we aimed at exploring the utility of genome sequences to devise a virus classification objectively, consistently, and fully. To this end, we have developed an approach for partitioning the genetic diversity of a virus family within a hierarchically organized framework. The developed approach provides quantitative support for both the delineated classification levels and the inferred taxa by devising the number and values of thresholds on intragroup genetic divergence at each level in a rational and family-wide manner. We named it DEmARC, which stands

for “DivErsity pArtitioning by hieRarchical Clustering” and refers to the English word “demarcation.” We extensively tested DEmARC on the proteome of the *Picornaviridae* (29), one of the most diverse and well-studied RNA virus families (23, 59) with numerous species that has been developed by one of the most active SGs (36, 37, 64). The picornavirus genome is a single-stranded positive-sense RNA (ssRNA+) with a single open reading frame that encodes a polyprotein (46, 50) flanked by two untranslated regions, 5′-UTR and 3′-UTR (69). The consistency and stability of the obtained results were evaluated by analyzing various data set derivatives which were compiled by varying the amount and/or the diversity of the input data, the alignment construction method, the measure of pairwise similarity, or a combination of parameters. In an accompanying paper (39), we analyze implications of the developed genetics-based classification for fundamental and applied research, through its comparison with virus phylogeny and taxonomy.

MATERIALS AND METHODS

Virus sequences and multiple alignments. Complete genome sequences for 1,234 picornaviruses available on 15 April 2010 at the National Center for Biotechnology Information GenBank/RefSeq (7) databases were downloaded using HAYGENS (61) into the Viroal platform (30). A multiple-amino-acid alignment of the polyproteins was produced using the Muscle program version 3.52 (21), and poorly conserved columns

TABLE 1 Composition of evaluation datasets

Data set ^a	Source of variation ^b				Virus diversity ^c	Domain diversity ^d	Date ^e	No. of sequences	No. of aa positions
	Region selection	Sequence sampling	Alignment building	Distance calculation					
E-Blocks	+	—	—	—	Picornia	1BCD, 2C, 3CD	2010	1,234	1,543
E-Capsid	+	—	—	—	Picornia	1BCD	2010	1,234	1,246
E-G1	+	+	—	—	Entero, sapelo	P1, 2BC, P3	2010	706	2,322
E-G2	+	+	—	—	Avihepato	P1, P2, P3	2010	65	2,255
E-G3	+	+	—	—	Hepato, tremo	P1, 2BC, P3	2010	58	2,060
E-G4	+	+	—	—	Parecho	P1, 2A6BC, P3	2010	44	2,302
E-G5	+	+	—	—	Kobu, sali*	P1, P2, P3	2010	12	2,401
E-G6	+	+	—	—	Aphtho	L, P1, P2, 3AB2CD	2010	267	2,480
E-G7	+	+	—	—	Cardio, seneca	P1, 2A4BC, P3	2010	39	2,219
E-G8	+	+	—	—	Tescho	L, P1, P2, P3	2010	31	2,242
E-G9	+	+	—	—	Cosa*	P1, P2, P3	2010	9	2,154
E-G10	+	+	—	—	Avihepato, parecho, aquama*	P1, 2A4BC, 3AB2CD	2010	110	2,312
E-G11	+	+	—	—	Aphtho, erbo, cardio, seneca, cosa*, tescho	P1, 2A4BC, 3AB2CD	2010	348	2,748
E-2008	—	+	—	—	Picornia	1BCD, 2C, 3CD	2008	685	2,374
E-2006	—	+	—	—	Picornia	1BCD, 2C, 3CD	2006	427	2,280
E-2004	—	+	—	—	Picornia	1BCD, 2C, 3CD	2004	181	2,269
E-Muscle	—	—	+	—	Picornia	1BCD, 2C, 3CD	2010	1,234	2,592
E-Clustal	—	—	+	—	Picornia	1BCD, 2C, 3CD	2010	1,234	2,269
E-PUD	—	—	—	+	Picornia	1BCD, 2C, 3CD	2010	1,234	2,446
E-PASC	+	—	+	+	Picornia	Complete genomes	2010	1,234	— ^f

^a For details on evaluation datasets, see Materials and Methods.

^b It is indicated which of four major variation parameters are affected with respect to the main data set, including 1,234 sequences and 2,446 positions.

^c Shown are abbreviated family or genera names; provisional or currently not recognized genera are marked with asterisks.

^d P1, P2, and P3 comprise capsid proteins (1A to 1D), nonstructural part 1 (2A to 2C), nonstructural part 2 (3A to 3D), respectively; for 2A and 3B designations, see reference 29.

^e Sampling date of data set according to GenBank annotation.

^f Not available due to the use of pairwise nucleotide alignments.

were further manually refined. The alignment construction was constrained by domain borders, most of which were delimited by known and predicted cleavage sites that are recognized by viral proteases in the polyprotein (46).

Data sets. In our study we used several data sets that are described below. Each data set has four characteristics: viruses, protein or genome region, alignment, and pairwise distances. We produced a family-wide data set that was treated as the main data set (M-2010) for the purpose of this study. It included regions of the polyprotein-wide alignment covering the family-wide conserved capsid proteins 1B, 1C, 1D (also known as VP2, VP3, and VP1, respectively) and the nonstructural proteins 2C, 3C, and 3D of 1,234 picornaviruses. Other genome regions were excluded from M-2010 due to the following reasons: (i) a protein is not conserved across the family (L*, L, 1A, 2A), (ii) a genome region was implicated in inter-species recombination (5'-UTR, 1A, 2A), or (iii) no confident alignment was obtained due to poor sequence conservation (2B, 3A, 3B, 3'-UTR). After discarding alignment columns that contained incomplete, termination, or nonspecified codons in one or more underlying nucleotide sequences, a final alignment of 2,446-amino-acid (aa) positions was derived. It was used to calculate pairwise evolutionary distances (PEDs) (see below) between all virus pairs.

To test the consistency and stability of results obtained for the main data set, in total 20 derivatives of M-2010, to which we refer as evaluation data sets (Table 1), were compiled by modulating one or several of the following four parameters: (i) genome region(s) selected for analysis, (ii) virus sequence sampling, (iii) alignment construction method, and (iv) measure of genetic divergence.

First, we extracted and concatenated blocks from the M-2010 alignment (with a lower limit of five and no upper limit on block width) that represent most informative alignment regions (evaluation data set E-Blocks) using BAGG (4, 12, 65). These blocks constitute regions of highest alignment quality/accuracy and account for ~63% alignment po-

sitions of the main data set. Second, we produced an M-2010 alignment derivative that included only the three capsid proteins (E-Capsid; ~51% alignment positions of the main data set). Third, 11 derivatives of the M-2010 alignment differing in respect to selection of viruses and/or proteins were compiled (E-G1 to E-G11). They represent either genus-like clusters or monophyletic sets of clusters (according to the phylogenetic analysis of M-2010) that include all domains conserved in the respective viruses of a data set. Fourth, three derivatives of the M-2010 alignment accounting for picornavirus sequences sampled up to a certain date were derived. The sampling dates used were 2, 4, and 6 years back in time and comprised, respectively, 685 (56% of sequences of the main data set; E-2008), 427 (35%; E-2006), and 181 (15%; E-2004) sequences. Fifth, we compiled two derivatives of the M-2010 alignment in which all protein domains were separately realigned without manual refinement using either the Muscle version 3.52 (21) or ClustalW version 2.0.12 (66) program (E-Muscle and E-Clustal, respectively). For all the evaluation alignments mentioned above PEDs were estimated. Sixth, we calculated pairwise uncorrected distances (PUDs) on the M-2010 alignment (E-PUD). Seventh, we calculated PUDs using all pairwise, genome-wide nucleotide alignments to emulate the PASC approach (E-PASC).

Estimation of pairwise distances. The metric used for classification is a measure of distance assigned to virus pairs, which was calculated based on a multiple-amino-acid alignment of respective virus sequences. To correct for multiple substitutions at the same sequence position, PED values were estimated by applying an maximum likelihood (ML) approach as implemented in the Tree-Puzzle program version 5.2 (57). The WAG amino acid substitution matrix (68) was used. PED values were compiled for the main and all but two evaluation data sets and analyzed in the same framework outlined below. For E-PUD and E-PASC data sets, PUDs were calculated. We note that any other type of pairwise distance measure could be utilized in the proposed framework as well. Conse-

FIG 2 Optimal bin number for the picornavirus-wide pairwise distance distribution. Shown is the χ^2 goodness-of-fit measure for approximating the picornavirus-wide PED distribution with normal probability densities using different bin sizes. Ten to 1,000 bins were tested, and the measure was normalized to a common scale of (0, 1). In the main analysis, a bin size of 0.01 (gray line) was used, which resulted in a significant fit with a χ^2 of 7.38 under a critical value of 117.0 with $n - p - 1 = 155$ degrees of freedom, $\alpha = 0.01$.

quently and unless otherwise stated, procedures utilizing PEDs that are described below were also applied to PUDs in this study. For brevity, PUDs will be mentioned only in places where the PUD and PED utilizations differ.

The DEMARC approach in a nutshell. We have developed a computational procedure for hierarchical classification of a set of viruses based on their PED values. A hierarchical classification is characterized by two major properties: (i) a number of levels that define the hierarchy and (ii) a number of clusters at each level that group the viruses unambiguously. These two characteristics are addressed by two steps in the developed procedure. At the first stage, the number of and support for levels in the hierarchical classification are determined by locating regions of discontinuity in the frequency distribution of PED values between all possible virus pairs. This is done by partitioning the distribution using a mixture of probability density functions. At the second stage, for each classification level a distance threshold within the respective region of discontinuity is identified. Such a threshold represents an upper limit on intragroup genetic divergence (measured by PEDs) at a level below which a virus pair is classified within the same cluster of that level. In the next two sections, the two stages of the procedure are explained in more detail.

DEMARC stage 1: locating regions of discontinuity in the pairwise distance distribution that define levels of a classification hierarchy. To identify regions of discontinuity in a PED distribution, we fitted a normal mixture model to the data. The fitted mixture model was subsequently used to assign a probability to each unique PED score that it originated from the underlying PED distribution. Consecutive PEDs with sufficiently low probabilities define a candidate region of distance discontinuity. The fitting (see below) was optimized by evaluating different bin sizes. For the M-2010 data set, the fit fluctuated sharply for large bin sizes and gradually converged to a steady state for bin sizes of <0.03 (Fig. 2). We used a bin size of 0.01 in all analyses.

To fit the mixture model, we first determined peaks in the PED distribution as positions with a frequency higher than those of the two adjacent PEDs. The entire PED distribution was then approximated by simultaneously fitting weighted probability densities to all determined peaks as well as to the background (noise). To do so, we utilized an expectation maximization (EM) approach adopted from reference 17 with the following three modifications: (i) normal instead of log-normal distributions were used, (ii) all peak components of the mixture were allowed to have separate variances, and (iii) the background component was modeled via a uniform distribution only. The normal mixture model (M) is defined by

$$M(d) = \sum_{k=1}^K w_k f_k(d) \quad (1)$$

with f_k being the probability density function that approximates component k for $(k = 1, \dots, K - 1)$ -determined peak components and the background component, component weights $w_1, \dots, w_K$ (such that they sum to

1), and pairwise distance d . The parameters of the distribution functions and the weights are estimated from the data by EM.

The deviation of the normal mixture model from the data was assessed using the following formula:

$$\chi^2 = \sum_{i=1}^b \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

with O_i and E_i being the observed and estimated frequencies (densities), respectively, of distance values d_i , and b being the number of histogram bins. It was compared to the critical value of the chi-square distribution with $n - p - 1$ degrees of freedom at a confidence level of 0.01 for n discrete distances and $P = 3 \cdot (K - 1) + 1$ estimated parameters (mean, variance, and weight of each peak component plus weight of the background component). The fit was significant for all data sets ($\alpha = 0.01$).

The goodness-of-fit (GOF) of the mixture model to the data was assessed using the following formula:

FIG 3 Corrected versus uncorrected picornavirus-wide pairwise distances. Plotted is corrected pairwise evolutionary distance (PED) versus pair-wise uncorrected distance (PUD) for the M-2010 data set. For intermediate and large distances, a saturation of PUD values is observed, as they do not account for the total amount of evolutionary work happened, e.g., for multiple substitutions at the same sequence position. Points on the dashed line (diagonal) have equal PED and PUD values.

FIG 4 Picornavirus-wide pairwise distance distribution and distance thresholds for partitioning. (A) Frequency distribution of ~760,000 PED values is shown for the M-2010 data set. In a first stage (see inset), peaks in the distribution were approximated using a mixture of normal distributions (red curves) together with an estimation of noise (purple horizontal line), with a goodness-of-fit of 0.972 (see Materials and Methods). For discrete distances along the distance range, TSM values (green bins) are shown. This measure is proportional to the probability of a particular distance not to be originated from one of the peak distributions. Consecutive distances with high TSM values provide candidate regions of distance discontinuity which can be used for partitioning the distribution and to infer levels of the hierarchical classification. In a second stage (B to D, top), distance threshold candidates within each region of discontinuity were probed in order to identify the threshold that minimizes the cumulative disagreement, the clustering cost (CC), of the potential clusters to the threshold. The change in the number of inferred clusters during this optimization is shown (B to D, bottom). The PED with the highest TSM score may differ from that with optimal CC (dashed vertical lines and arrows in blue). For the four top-ranked thresholds (including the trivial one at maximum distance), the number of inferred clusters is indicated above the black horizontal bars in A. The bars delimit respective intragroup distance ranges. The pairwise distance scale reflects the estimated number of amino acid substitutions per site on average.

$$\text{GOF} = 1 - \frac{X^2}{b} \tag{3}$$

After fitting, a threshold support measure (TSM) was compiled for each (unique) PED value according to the following formula:

$$\text{TSM}(d) = -\log_{10} \sum_{k=1}^{K-1} 2w_k \min\{F_k(d), 1 - F_k(d)\} \tag{4}$$

with F_k being the value of the cumulative distribution function for peak component k . Due to the nature of the normal cumulative function, which has a value of 0.5 at the distribution mean, we introduced the factor 2 to ensure that the TSM theoretically can be 0 at the lowest point. Peaks in the TSM distribution were used to define candidate regions of distance discontinuity, which were ranked according to their TSM values. The top-ranked candidates define the levels of a classification hierarchy.

DEmARC stage 2: identification of distance thresholds that delimit

level boundaries. At the second stage, we sought to determine a distance threshold for each classification level. To this end, all PEDs inside the respective region of distance discontinuity (between adjacent local minima in the TSM distribution; see above) were probed. For each probed threshold, single linkage clustering (SLC) was applied to group viruses into clusters. According to SLC, each virus is separated from at least one other virus in the cluster by a distance that is below the applied distance threshold. Consequently, some PEDs may exceed the threshold, collectively referred to as violating PEDs. The total extent of such violations across all clusters was summarized to define a cost for the probed distance threshold. This so-called clustering cost (CC) was calculated as follows:

$$\text{CC} = \sum_{c=1}^C \sum_{d_c > t} \left(\frac{d_c - t}{t} \right) \tag{5}$$

for inferred clusters $c = 1, \dots, C$, intragroup distance values d_c , and dis-

TABLE 2 Quality of classification levels and accordance of classifications built for the main (M-2010) and evaluation (E-x) datasets

Data set ^a	Species			Genus			Supergenous		
	No.	CC ^b	CA ^c	No.	CC ^b	CA ^c	No.	CC ^b	CA ^c
M-2010	38	5.54	1	16	0.16	1	11	0.20	1
E-Blocks	39	0.70	0.925	16	0.71	1	10	0	0.750
E-Capsid	37	4.82	0.630	14	0	0.667	9	12.78	0.667
E-G1 ^d	17	1.63	1	2	0	1	1	—	—
E-G2 ^d	1	—	—	1	—	—	1	—	—
E-G3 ^d	2	0	1	2	—	—	1	—	—
E-G4 ^d	2	0	1	1	—	—	1	—	—
E-G5 ^d	3	0	1	2	0	1	1	—	—
E-G6 ^d	3	0	1	1	—	—	1	—	—
E-G7 ^d	3	0	1	2	0	1	1	—	—
E-G8 ^d	1	—	—	1	—	—	1	—	—
E-G9 ^d	4	0	1	1	—	—	1	—	—
E-G10 ^d	4	0	1	3	0	1	3	—	—
E-G11 ^d	12	0	1	7	0	1	5	0	1
E-2008 ^d	24	0.13	0.885	12	0	1	10	0	1
E-2006 ^d	18	0	1	9	0	1	7	0	1
E-2004 ^d	16	0	1	8	0	1	7	0	1
E-Muscle	39	7.65	0.925	16	0.41	1	11	0	1
E-Clustal	39	5.43	0.925	16	0.03	1	11	0	1
E-PASC	36	3.82	0.762	16	23.16	0.684	0	—	0
E-PUD	38	6.40	1	16	3.06	1	10	0.27	0.750

^a See Table 1 for details on evaluation datasets.

^b Shown is the clustering cost (CC) representing the cumulative disagreement of all clusters at a level; a value of 0 represents absolute (optimal) agreement due to perfect separation of all clusters (see Materials and Methods for details).

^c Shown is a clustering accordance (CA) value of a classification relative to the main data set; a value of 1 represents identical classifications (see Materials and Methods for details).

^d This data set has only a fraction of viruses presented in M-2010. Consequently, CA values reflect the agreement between two datasets in respect to this virus subset.

—, not shown for trivial clusterings formed by a single taxon.

tance threshold value t . The CC is a simplification of the modification cost defined in reference 70, the computation of which turned out to be prohibitively expensive for data set sizes of this study. In the ideal case, when there are no violating PEDs, CC is zero; otherwise, CC is >0 . For each classification level, the optimal threshold among all probed candidates was determined by selecting the one with minimum cost.

Quantification of the quality of clusters. For each cluster of an inferred classification, we quantified its quality as the fraction of intragroup pairwise distances not exceeding the distance threshold of the respective level, to which we refer as cluster quality (cq). A cluster is considered complete if the cq value was 1 and incomplete otherwise ($0 < cq < 1$).

Comparison of classifications. The classification for M-2010 was compared separately to those obtained for each evaluation data set at each inferred classification level. The fraction of matching clusters in the compared classifications was quantified using the following measure, to which we refer as clustering accordance (CA):

$$CA = \frac{X}{Y + X + Z} \quad (6)$$

with X being the number of common clusters (those with identical virus compositions) in the two classifications, and Y and Z the number of clusters which are unique to the classification for M-2010 and an evaluation data set, respectively. In each comparison, only the subset of viruses common to both data sets was considered. Identical classifications result in CA values of 1; otherwise, CA is <1 .

Implementation details. The DEMARC framework was implemented using custom Perl (51) and R (52) scripts. A complete analysis of the M-2010 data set, excluding alignment building, took about 4 h 30 min on a Linux machine with 4 central processing units (CPUs), 2660 MHz, and 4 GB RAM.

RESULTS

GENETIC classification of picornaviruses: distance measure, levels, and thresholds. Using an ML approach, PED values were compiled for all pairs of the 1,234 picornavirus sequences in the main alignment data set M-2010 (n , ~760,000). These distances are evolutionary based (an evolutionary model is involved in the calculation) and corrected for multiple substitutions at the same sequence position. An effect of this correction is already evident at distances above 0.1 in a steadily growing deviation from the linear relation between PED and PUD distributions calculated for this data set (Fig. 3). When PUDs approach ~0.8, PEDs already reach ~2.2, outpacing the former by more than an order of magnitude at this and greater divergence. A PED frequency distribution is multimodal, revealing a number of peaks separated by areas of low frequency in the pairwise distance range of 0 to 2.78 (in units of average number of substitutions per site) (Fig. 4A). Peaks correspond to dominant distances among various virus pairs, and their heights are affected by virus sampling bias. Consequently, peaks in the distribution should not be discarded solely based on their relatively minor size/height.

By fitting a normal mixture model to the picornavirus PED distribution and calculating TSM values along the PED range (see Materials and Methods), three most strongly supported regions of discontinuity were identified (Fig. 4). The highest TSM was assigned to the region at the intermediate distance of around 1.2 (TSM of 76.1), followed by the ones at the low distance of 0.43 (39.0) and the intermediate distance of 0.93 (14.2) (Fig. 4A). The next best region, not considered in this study, had a substantially lower support with a TSM of 6.5.

FIG 5 Impact of weakly conserved alignment regions and selection of capsid proteins on the GENETIC classification. Frequency distributions of ~760,000 PED values formed by 1,234 picornaviruses are shown for the following evaluation data sets: a data set containing only highly conserved alignment regions (blocks) of the main data set (A), and a data set containing only the three capsid proteins 1B, 1C, and 1D (B). The goodness-of-fit values are 0.987 and 0.992, respectively. For details see Materials and Methods and Fig. 4.

Next, we sought to identify an optimal distance threshold within each of the three regions of discontinuity determined above. To this end, PEDs within a region were probed as potential distance thresholds, and a cost was assigned to each of them using the CC measure (see Materials and Methods). This cost function showed multiple local minima within a region of discontinuity, each following a change in the underlying number of clusters (Fig. 4B to D). The candidate with the minimal cost was selected as the optimal threshold of a region, although we noted that the cost value of the next best candidate could be only slightly worse. We found that in the three regions of discontinuity the PEDs with optimal CC values do not match those with highest TMS values but rather are located in their vicinity (Fig. 4B to D; Table 2). The optimal thresholds (in the order from left to right in the PED distribution) and the number of clusters they determine were as follows: 0.37 (38 clusters), 0.905 (16), and 1.161 (11) (Fig. 4B to D). By applying these three thresholds to the picornavirus genetic diversity, we derive a hierarchical classification with three levels (species, genus, and supergenus) which we refer to as the “GENETIC classification” (Fig. 4A) (39).

Consistency and stability of the GENETIC classification. Using the CC and CA measures (see Materials and Methods), we proceeded to evaluate the consistency and stability of the GENETIC classification by analyzing 20 alignment derivatives which were produced by varying the amount and/or diversity of the input data, the alignment construction method, the measure of pairwise similarity, or a combination of two parameters. In many instances, we observed high quality (CC equal or close to zero), while agreement varied considerably ($0 \leq CA \leq 1$) (Table 2).

In the first evaluation test, we analyzed a possible impact of weakly conserved protein residues on the virus classification. To this end, protein residues that formed ~37% of the alignment

TABLE 3 Differences in classifications built for the main (M-2010) and evaluation (E-x) datasets

Difference		Assignment to taxon ^a in data set ^b							
Virus(es) involved	At level	M-2010	E-Blocks	E-Capsid	E-2008	E-Muscle	E-Clustal	E-PASC	E-PUD
Bovine kobuviruses	Species	BKoV	BKoV α	BKoV α	—	BKoV α	BKoV α	BKoV α	—
Porcine kobuviruses	Species	BKoV	BKoV β	BKoV β	—	BKoV β	BKoV β	BKoV β	—
HRV-C 026, NY-074, NAT001, QPM	Species	HRV-C α	—	HRV-C	HRV-C $\alpha\beta$	—	—	HRV-C	—
HRV-C 025	Species	HRV-C β	—	HRV-C	HRV-C $\alpha\beta$	—	—	HRV-C	—
HRV-C N4, N10, NAT045	Species	HRV-C γ	—	HRV-C	—	—	—	HRV-C	—
HRV-A	Species	HRV-A α	—	HRV-A	—	—	—	HRV-A	—
HRV VR-1118, VR-1155, VR-1301	Species	HRV-A β	—	HRV-A	—	—	—	HRV-A	—
HEV-A	Species	HEV-A	—	HEV-A	—	—	—	—	—
Baboon enterovirus A13	Species	SiEV-B	—	HEV-A	—	—	—	—	—
FMDV type Asia 1, A, O, C	Species	FMDV	—	FMDV α	—	—	—	—	—
FMDV type SAT1, SAT2, SAT3	Species	FMDV	—	FMDV β	—	—	—	—	—
Avian sapeloviruses	Genus	Sa	—	—	—	—	—	Sa α	—
Porcine and simian sapeloviruses	Genus	Sa	—	—	—	—	—	Sa β	—
Kobuviruses	Genus	Ko	—	KoSK	—	—	—	KoSK	—
Sali- and klasseviruses	Genus	SK	—	KoSK	—	—	—	KoSK	—
Hepatoviruses	Genus	He	—	HeTr	—	—	—	—	—
Tremoviruses	Genus	Tr	—	HeTr	—	—	—	—	—
Cardio- and senecaviruses	Supergen	CaSe	CaSeCo	CaSeCoEr	—	—	—	—	CaSeCo
Cosaviruses	Supergen	Co	CaSeCo	CaSeCoEr	—	—	—	—	CaSeCo
Erboviruses	Supergen	Er	—	CaSeCoEr	—	—	—	—	—
Picornaviruses	Supergen	$n = 11$	$n = 10$	$n = 9$	—	—	—	None	$n = 10$

^a Abbreviations: Sa, sapeloviruses; Ko, kobuviruses; SK, sali- and klasseviruses; Ca, cardioviruses; Se, senecaviruses; Co, cosaviruses; Er, erboviruses. A dash denotes that an evaluation data set is in accordance with the main data set.

^b See Table 1 for details on evaluation datasets.

FIG 6 Reproducibility of the GENETIC classification on the species level, part one. Frequency distributions of PED values are shown for supergenera G1 to G5 of the main data set (A to E) or a combination of three supergenera (F). PED values were compiled based on alignments covering all cluster-wide conserved domains (Table 1). Viruses currently not recognized by the ICTV are marked with asterisks. (E) An alternative threshold is indicated which would result in four instead of three species clusters (dashed line and names). The goodness of fit is in the range from 0.751 to 0.965. For details, see Materials and Methods and Fig. 4.

columns in M-2010 with the lowest conservation scores (4, 12) were removed from the analysis (E-Blocks data set) (Table 1; Fig. 5A). Compared to M-2010, the E-Blocks classification showed one difference on the species level ($CA = 0.925$): recently discovered porcine kobuviruses formed a species separate from *Bovine kobuvirus*. On the genus level, perfect agreement between the two classifications ($CA = 1$) was observed, while on the supergenus level an expansion of the *Cardiovirus*/*Senecavirus* supergenus with cosaviruses was evident ($CA = 0.750$) (Tables 2 and 3). For both levels at which a disagreement was observed, E-Blocks outranked M-2010 in respect to the classification quality by CC: 0.70 versus 5.54 (species) and 0 versus 0.20 (supergen), respectively.

In the second evaluation test, we analyzed the dependence of the classification on the choice of proteins. We compared results for M-2010 with those obtained for a data set using the three main capsid proteins (1BCD; E-Capsid), which are often regarded as

representing picornaviruses. An outstanding support ($TSM = 19.3$, $CC = 0$) was observed only for the genus level, while these values for species (5.7, 4.82) and supergenus (5.5, 12.78) levels were considerably worse, and they were on par with the support value (8.7, 7.4) for another level below species (Fig. 5B; Table 2). The classification produced for E-Capsid differed from the M-2010 classification in a number of aspects and showed the lowest agreement among all PED-based evaluation data sets. On the species level, several clusters from different genera were affected ($CA = 0.630$). They include *Human rhinovirus A* (HRV-A; accepted otherwise separated HRV-A β), *Human rhinovirus C* (HRV-C; one instead of three clusters), *Foot-and-mouth disease virus* (FMDV; split into two), porcine/bovine kobuviruses (split into two), and *Human enterovirus A* (accepted a virus that was otherwise classified with *simian enterovirus B* [SiEV-B]). At the genus level, 14 instead of 16 genera were observed ($CA = 0.667$):

FIG 7 Reproducibility of the GENETIC classification on the species level, part two. Frequency distributions of PED values are shown for supergenera G6 to G9 of the main data set (A to D) or a combination of five supergenera (E). PED values were compiled based on alignments covering all cluster-wide conserved domains (Table 1). Viruses currently not recognized by the ICTV are marked with asterisks. (D) No fitting of probability densities could be obtained due to an insufficient number of sequences ($n = 9$). The goodness of fit is in the range from 0.751 to 0.965. For details, see Materials and Methods and Fig. 4.

Hepatitis virus and *Tremovirus* (47) as well as *Kobuvirus* and *salivirus* (43), respectively, were united. At the supergenus level, 9 rather than 11 clusters were identified ($CA = 0.667$): the supergenus *Cardiovirus/Senecavirus* was expanded by the inclusion of *Erbovirus* and cosaviruses (Tables 2 and 3).

In the third evaluation test, we analyzed a combined impact of protein selection and sequence diversity on the virus classification. To this end, we scrutinized 11 virus data sets that were formed by viruses representing supergenus clusters according to the M-2010 classification (from E-G1 to E-G9) or monophyletic clades comprising several (super)genera (E-G10, 4 species; E-G11, 12 species) (Table 1; Fig. 6 and 7). For each of these 11 data sets, all cluster-wide conserved domains were included in the respective alignments. E-G1, for instance, includes the same set of entero- and sapeloviruses found in M-2010, but the two data sets differ con-

siderably in terms of protein composition. Species classifications obtained for each of the analyzed evaluation data sets perfectly matched ($CA = 1$) that of M-2010 (Table 2).

In the fourth evaluation test, we analyzed the dependence of the GENETIC classification on sequence sampling by analyzing virus data sets available at three time points in the past: the years 2008 (E-2008), 2006 (E-2006), and 2004 (E-2004) (Table 1; Fig. 8). Together with M-2010, these data sets encompass a variation in virus sampling in the range of 181 to 1,234 sequences that was analyzed in this study. On the genus and supergenus levels, perfect agreement among classifications for M-2010 and the three evaluation data sets was observed. Naturally, these comparisons involved only a subset of viruses of M-2010 that was available at a specific time point in the past. At the species level, only a single difference was evident: for E-2008, the clusters HRV-Cα and

FIG 8 Impact of virus sampling on the GENETIC classification. Frequency distributions of PED values are shown for evaluation data sets formed by picornaviruses sampled until 2 years (A), 4 years (B), and 6 years (C) ago with respect to the sampling time of the main data set. The goodness-of-fit values are 0.973, 0.978, and 0.953, respectively. For details, see Materials and Methods and Fig. 4.

HRV-C β were united ($CA = 0.885$) (Tables 2 and 3), resulting in two instead of three (for M-2010) species-like clusters for viruses jointly classified as *Human rhinovirus C* in the current taxonomy.

In the fifth evaluation test, we assessed an impact of alignment construction on the virus classification, using the E-Muscle and E-Clustal evaluation data sets (Table 1; Fig. 9A, B). The GENETIC classification of both evaluation data sets matched that of M-2010 on the genus and supergenus levels and showed a single common deviation at the species level ($CA = 0.925$), which involved bovine and porcine kobuviruses, a mismatch already observed for E-Blocks (Tables 2 and 3).

In the sixth evaluation test, we analyzed the impact of the sole choice of distance measure, PED (M-2010) versus PUD (E-PUD), on the GENETIC classification (Fig. 9C). The only difference was that the supergenus *Cardiovirus/Senecavirus* merged with

the genus formed by cosaviruses for E-PUD ($CA = 0.750$) (Tables 2 and 3).

In the seventh evaluation test, we compiled pairwise, genome-wide nucleotide alignments to calculate PUDs in order to emulate the PASC application (6), the standard tool at NCBI. A classification for the resulting data set, E-PASC, was derived (Fig. 9D) by using DEmARC. Its comparison to that of M-2010 reveals most drastic differences. On the species level ($CA = 0.762$), the E-PASC classification has *Human rhinovirus A* and HRV-A β united, *Human rhinovirus C* viruses forming a single cluster, and porcine kobuviruses forming a cluster separate from *Bovine kobuvirus*. On the genus level ($CA = 0.684$), the avian sapelovirus formed a cluster separate from other sapeloviruses and saliviruses joined with *Kobuvirus*, which are recognized as a supergenus cluster in the M-2010 classification. Furthermore, the supergenus level was not recovered in the E-PASC classification ($CA = 0$). Each of the above deviations concerns clusters whose median or extreme PED value is in the immediate vicinity of a threshold in the M-2010 classification (data not shown), indicating that the recovery of such clusters is most sensitive to the choice of key parameters, the default values of which differ between PASC and DEmARC.

Accommodation of virus sampling bias by the GENETIC classification. It is generally acknowledged that the current sampling of the picornavirus diversity is limited and biased (29, 41). This variation is illustrated spectacularly for viruses of the M-2010 data set: 82% of the least populated species account for only 18% of the viral genomes (Fig. 10A). The lack of correlation between the sampling size and the cluster completeness of species attests to the tolerance of the GENETIC classification to this variation. The sampling unevenness is also evident when calculating the skewness on the distribution of number of sequences per cluster at each level. (Skewness is a measure of asymmetry of a distribution which is positive or negative when a distribution is right-tailed or left-tailed, respectively, and zero when it is symmetric). It was 2.51 (for species), 2.99 (genera), and 2.32 (supergenera). In contrast, the unevenness of frequency distributions of taxa at higher classification levels—species among genera and genera among supergenera—is progressively diminishing (Fig. 10B and C).

DISCUSSION

Here we present a quantitative, evolutionary-based framework (DEmARC) for computational partitioning of the genetic diversity of a virus family with the dual goal of revealing its internal structure and building a rational genetics-based virus classification. Applying DEmARC to hundreds of genome sequences of picornaviruses, we produced the GENETIC classification of the family *Picornaviridae* that was largely tolerant to the choice of virus sampling and alignment construction method, parameters that are of particular importance for taxonomy. In the accompanying paper (see reference 39), we show that this classification closely approximates the expert-based taxonomy of the family, while providing a basis for biological interpretations not available in the current taxonomy framework.

DEmARC framework: choices, novelties, and challenges. Below we discuss choices, novelties and challenges of the DEmARC framework (Table 4) that concern (i) input data, (ii) alignment-building procedure, (iii) measure of genetic divergence, (iv) decision-making in virus clustering, and (v) classification robustness.

(i) **Input data.** We chose amino acid over nucleotide se-

FIG 9 Impact of alignment construction and incorporation of PASC elements into the DEMARC framework on the GENETIC classification. Frequency distributions of ~760,000 PED or PUD values formed by 1,234 picornaviruses are shown for the following evaluation data sets: PEDs were calculated using the main data set that was automatically realigned without manual intervention using Muscle (A) and ClustalW (B), PUDs were calculated using the main data set (C), and PASC-based genome-wide PUDs were calculated (D). The goodness-of-fit values are 0.982, 0.993, 0.865, and 0.956, respectively. For details, see Materials and Methods and Fig. 4.

quences, since proteins accept fewer replacements than polynucleotide sequences, which is of particular importance in analyses of RNA viruses, including picornaviruses, due to their extraordinarily high mutation rates (19, 20, 33, 54). In the picornavirus protein data set, viruses are separated by PEDs that already amount up to 2.8 replacements per position on average in the conserved proteins.

We were interested to include as many genome positions as possible in the analysis upon reasoning that the more genome positions, the more authentic an obtained classification. Since many positions contribute to the classification, expanding their number in a data set may moderate or even negate effects caused by across site rate variation due to mutation and local recombination. Technically, the choice is limited to orthologous genes and their products that are known to diverge by vertical descent in the entire data set. In our study, we analyzed all orthologous proteins conserved across all viruses in the data set (29). They account for ~80% of the entire picornavirus proteome for M-2010 and even larger shares in the evaluation data sets E-G1 to E-G11. This approach can be contrasted with a practice to restrict the analysis to single gene/protein, e.g., references 1, 8, and 44; commonly a structural protein is used (see, for example, references 3, 6, 44, 49, 60, and 71). In our study, we observed that the number of clusters in M-2010 compared to E-Capsid was somewhat larger for all three levels (Table 2). This observation indicates that phylogenetic signals in the conserved capsid and replicase proteins can produce a cumulative effect, further supporting their combined use in the picornavirus taxonomy (37).

To produce a GENETIC classification, we typically utilized

PED values that are products of an evolutionary inference. In evolutionary analyses involving multiple genes or proteins (like in this study), it is common to evaluate and exclude the contribution of recombination. If a portion of a genome has originated through recombination while another part evolved only by mutation, evolutionary inferences for the combined data set would be biologically misleading. Unfortunately, the scale of recombination in our data set (M-2010) remained uncharacterized, since its size (1,234 sequences, 2,446 alignment positions) is too large to apply available tools for the identification of recombination events (24, 38, 45). Nevertheless, there are several reasons to believe that recombination was limited and accommodated by the classification. We excluded from our analysis (M-2010) three regions, 5'-UTR, VP4, and 2A, for which interspecies recombination has been reported (55, 62). Outside these regions, recombination was reported exclusively for closely related viruses of the same species, although few viruses were characterized in this respect (5, 9, 10, 34, 40–42, 53, 55, 62, 63, 72). These intraspecies recombination events are not expected to be detrimental for our analysis since, following taxonomy, it was not concerned with virus clustering below the species level. Furthermore, the GENETIC classification does recover the ICTV-defined species structure, with most species obeying the family-wide limit on intragroup genetic divergence (Fig. 10A) (39). The latter observation implies that recombination between viruses of M-2010 must be restricted within the species boundaries for genes encoding the most conserved proteins. This notion is further supported by the excellent agreement between classifications obtained for M-2010 and the evaluation data sets (E-G1 and E-G10) representing different subsets of the family

FIG 10 Sampling size of taxa and completeness of species in the GENETIC classification. (A) Shown is a binary square matrix of 1,234 viruses derived from the M-2010 PED matrix. Virus pairs whose PED does not exceed the species distance threshold are shown as black dots that form 38 species-specific squares along the matrix diagonal; other pairs are in white. Viruses along both coordinates are grouped by species, and species are ordered by descending virus sampling size. Note that no black dots are observed outside the squares, which is expected in classifications by SLC. For the most-populated clusters, their names and the number of sampled sequences are shown. Zoom-ins and quality values (cq) which are <1 are provided in brackets for three species for which some PEDs (depicted as empty spaces within black squares) exceeded the threshold (incomplete clusters). For all other clusters, the cq value was 1. (B) Shown is a binary square matrix of 38 species that form 16 genus-specific squares along the matrix diagonal. Species pairs from the same genus are in black, others in white. Species along both coordinates are grouped by genus, and genera are ordered by descending species sampling size. All genera are shown as if they were complete, despite the fact that the cluster formed by *Enterovirus* has a cq of only 0.9998. For the most-populated clusters, their identity and the number of sampled species are indicated. (C) Shown is a binary square matrix of 16 genera that form 11 supergenus-specific squares along the matrix diagonal. Genus pairs from the same supergenus are in black, others in white. Genera along both coordinates are grouped by supergenus, and supergenera are ordered by descending genus sampling size. All supergenera are shown as if they were complete, despite the fact that the cluster formed by *Enterovirus/Sapelovirus* has a cq of only 0.9975. The number of sampled genera is indicated for the largest cluster.

(Table 2). Such an agreement is not expected if recombination acts across genus boundaries. Based on these observations, we conclude that any interspecies recombination in family-wide conserved proteins, if it had happened, must have been (very) limited and hence does not affect the reliability of the inferred classification.

(ii) Alignment-building procedure. We calculated pairwise distances based on a multiple alignment, which compared to widely used pairwise alignments (1, 2, 6, 16), is expected to improve the reconstruction accuracy of orthologous relationships of sequence residues (27, 31). Surprisingly, the choice of method for obtaining a multiple sequence alignment was not as critical as might be expected. The use of either Clustal- or Muscle-based alignments or a manually curated alignment, which were different in the number of gaps and their distribution (Table 1), had little impact on the GENETIC classification (Table 2), a finding which readily permits (automated) reproduction of results.

(iii) Measure of genetic divergence. We made use of a distance measure that is evolutionary based and corrects for multiple substitutions at the same sequence site. It can be calculated with publicly available tools, for instance Tree-Puzzle (57), as used in this study. Indeed, we observed a nonlinear relationship between PUD and PED values (Fig. 3). As a result, virus pairs that occupy the 2nd half in the PED distribution are found in the last 20% of the distribution of PUDs (Fig. 9C, compare results for M-2010 in Fig. 4 and E-PUD). This relative compression of large distances may result in a relatively lower resolution of distant relationships that could affect the delineation of higher-order taxonomic levels (subfamily for instance) in future analyses of this and other virus families. When PUDs were combined with the use of pairwise nucleotide alignments (E-PASC data set), the supergenus level was already not recoverable (Fig. 9D).

We note that it would be worth exploring the use of patristic distances instead of PEDs. The patristic distance between two viruses is defined as the amount of substitutions since they shared a common ancestor in evolution, and thus a phylogenetic tree is involved in its calculation. The reconstruction of such a tree in practice, using sophisticated methods (maximum likelihood or Bayesian), however, turned out to be computationally very expensive for the data sets analyzed in this study and hence was not pursued.

(iv) Decision-making in virus clustering. In prior virus classification studies, researchers were commonly concerned with the placement of demarcation criteria by following ranks and taxa already established by the respective ICTV study group. The framework developed in this study operates without following an *a priori*-defined number of taxonomic ranks, since it seeks to unravel the intrinsic hierarchical structure embodied in the data. This is achieved in a quantitative manner by searching for regions with strongest support for discontinuity in the PED distribution. The selection of these regions controls the number of levels and their hierarchy. We acknowledge, however, that this selection is yet to be placed in a statistical framework. For the picornavirus data set, the three top-ranked regions considerably outranked all other candidates (Fig. 4), making their selection relatively straightforward. This might not be the case for other virus families with relatively poor virus sampling, and it was observed upon the analysis of E-G5 (Fig. 6E). Additional research will be necessary to further improve this part of the framework.

The subsequent delineation of a demarcation threshold on in-

TABLE 4 Comparison of pairwise distance-based classification approaches of this study, the standard tool at NCBI, and other studies^a

Aspect	Parameter for indicated classification approach		
	DEmARC M-2010	PASC	Others
Genome regions included	All family-wide conserved proteins	Complete genomes	Single genes/proteins, their combinations
Sequence type	aa	nt	aa and/or nt
Alignment	Multiple	Pairwise	Multiple, pairwise
Distance measure	Corrected	Uncorrected	Uncorrected, corrected
No. of taxonomic levels	Data derived	<i>A priori</i> defined	<i>A priori</i> defined
Threshold determination ^b	Objective	Subjective	Subjective

^a Classifications are indicated as follows: DEmARC M-2010, this study; PASC, the standard tool at NCBI; others, other studies.
^b It is indicated how thresholds for partitioning of the distance distribution are determined: using either an objective, data-driven approach (objective) or other means, including rough, subjective placement and missing description (subjective).

tragroup genetic divergence at each classification level is done in a fully objective manner by locally restricted cost optimization. The approach seeks to minimize the global disagreement across all clusters at a level. We note that all violations (PEDs exceeding the distance threshold) are weighted equally independent of the size of the respective cluster or the number of other intragroup PEDs that obey the threshold. Future research is needed to scrutinize more sophisticated cost functions and their possible impact on decision-making. Besides, we chose to derive clusters using SLC, which implies that it is sufficient for a virus to be similar enough to a single other virus of a cluster in order to be classified within that cluster. From a biological perspective, this seems to be more meaningful than the opposite approach—complete linkage clustering (CLC)—where no intragroup violations are allowed but clusters may overlap (intergroup divergence below the distance threshold). Nevertheless, we tested the impact of CLC on M-2010 and found that it was small, with only one difference at the species level (clade C rhinoviruses are grouped in two instead of three clusters) and one at the supergenus level (cosaviruses are grouped with *Cardiovirus/Senecavirus*) (unpublished observation). Most importantly, however, in either case a consistent demarcation criterion is imposed on all clusters of a level regardless of the virus sampling sizes and diversities, parameters which strongly shape decision-making in traditional virus taxonomy. To our knowledge, the threshold identification in DEmARC presents the first application of a rigorous approach to the problem (Table 4).

(v) **Classification robustness.** One of the grand challenges in developing an objective classification of a virus family is the lack of a positive control that may serve as a gold standard. It could be argued that the expert-based ICTV taxonomy should be used as the ultimate standard, and we do compare the GENETIC classification with the taxonomy of picornaviruses (39). This comparison is informative and, if experts recognize merits of the GENETIC classification, it could prompt a revision of taxonomy. It is because of the prospect of such a revision that the picornavirus taxonomy may not be regarded as a scientifically valid gold standard for the GENETIC classification. In this context, we may not know how close the GENETIC classification is to its ultimate standard that remains unknown. Consequently, ranking alternative classifications by quality estimates using objective measures like CC remains the most practical way to evaluate the performance of the developed approach. In this study, we selected the M-2010-based classification as the standard using different considerations discussed elsewhere in this paper. However, we noticed that the E-Blocks-based classification outranked the M-2010-based one under the CC criterion at the two levels of

hierarchy at which they deviate (Table 2). The observed differences between these two classifications were only few and minor, all involving problematic virus clusters. This situation may change with the expansion of the number of genomes analyzed in the future, and alignments processed with BAGG, e.g., E-Blocks, may prove to be superior to those unprocessed, e.g., M-2010, in virus classification. This development would be in line with the acknowledged positive effect of purging multiple alignments from poorly conserved columns on phylogeny reconstruction (65). We also note that the E-Blocks-based classification shows considerable support for a fourth level of hierarchy above the supergenus level (Fig. 5A). This indicates that switching from unprocessed to block-based alignments could be associated with additional large-scale consequences for taxonomy.

General conclusions. During the last decade, genome sequences have emerged as the primary and principal characteristic for all known viruses. The flood of genome sequences overwhelmed the traditional decision process designed to classify viruses. We here have introduced a consistent and objective framework that addresses this challenge in a proof-of-principle study using the family *Picornaviridae*. We thereby follow a parallel development in taxonomic studies of cellular organisms where recent advancements are increasingly brought by the analysis of molecular data, jointly summarized under the label “DNA barcoding” (11, 32). The produced genome-based partitioning of the picornavirus genetic diversity could assist the ICTV in decision-making and be used to improve the connection between virus taxonomy and fundamental and applied research (39). Technically, DEmARC can be fed with partial genomes, the analysis of which may be valuable for taxonomy or other purposes, although this is yet to be explored. We started to seek benefits of the developed computational framework in analyses of other (RNA) virus families, and the DEmARC-mediated taxonomy of coronaviruses has recently been approved by ICTV (14).

ACKNOWLEDGMENTS

We are indebted to Johan Faase for his involvement in the initial phase of this project, Hans van Houwelingen for commenting on a manuscript draft, 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 Virology 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 program), and Leiden University Fund (Special Chair in Applied Bioinformatics in Virology).

REFERENCES

- Adams MJ, et al. 2004. The new plant virus family Flexiviridae and assessment of molecular criteria for species demarcation. *Arch. Virol.* 149: 1045–1060.
- Adams MJ, Antoniw JF, Fauquet CM. 2005. Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch. Virol.* 150: 459–479.
- Ando T, Noel JS, Fankhauser RL. 2000. Genetic classification of “Norwalk-like viruses.” *J. Infect. Dis.* 181:S336–S348.
- Antonov IV, Leontovich AM, Gorbalenya AE. 2008. BAGG - Blocks Accepting Gaps Generator, version 1.0. <http://www.genebee.msu.ru/~antonov/bagg/cgi/bagg.cgi>.
- Arita M, et al. 2005. A Sabin 3-derived poliovirus recombinant contained a sequence homologous with indigenous human enterovirus species C in the viral polymerase coding region. *J. Virol.* 79:12650–12657.
- Bao Y, Kapustin Y, Tatusova T. 2008. Virus classification by pairwise sequence comparison (PASC), p 342–348. *In* Mahy BWJ, Van Regenmortel MHV (ed), *Encyclopedia of virology*, vol 5. Elsevier, Oxford, United Kingdom.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2010. GenBank. *Nucleic Acids Res.* 38:D46–D51.
- Bernard HU, Burk RD, Chen ZG, van Doorslaer K, zur Hausen H, de Villiers EM. 2010. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401:70–79.
- Bessaud M, Joffret ML, Holmblat B, Razafindratsimandresy R, Delpeyroux F. 2011. Genetic relationship between cocirculating human enteroviruses species C. *PLoS One* 6:e24823.
- Brown B, Oberste MS, Maher K, Pallansch MA. 2003. Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region. *J. Virol.* 77:8973–8984.
- Casiraghi M, Labra M, Ferri E, Galimberti A, de Mattia F. 2010. DNA barcoding: a six-question tour to improve users' awareness about the method. *Brief. Bioinform.* 11:440–453.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552.
- Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. *Science* 312:1795–1798.
- de Groot RJ, et al. 2012. Family *Coronaviridae*, p 806–828. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Academic Press, Amsterdam, The Netherlands.
- Delwart EL. 2007. Viral metagenomics. *Rev. Med. Virol.* 17:115–131.
- de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. 2004. Classification of papillomaviruses. *Virology* 324:17–27.
- Dijkstra M, Roelofs H, Vonk RJ, Jansen RC. 2006. Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics* 6:5106–5116.
- Domingo E. 2007. Virus evolution, p 389–421. *In* Knipe DM, et al (ed), *Fields virology*. Wolters Kluwer, Lippincott Williams & Wilkins, Philadelphia, PA.
- Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. U. S. A.* 96:13910–13913.
- Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9:267–276.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Edwards RA, Rohwer F. 2005. Viral metagenomics. *Nat. Rev. Microbiol.* 3:504–510.
- Ehrenfeld E, Domingo E, Roos RP (ed). 2010. *The picornaviruses*. ASM Press, Washington, DC.
- Etherington GJ, Dicks J, Roberts IN. 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics* 21:278–281.
- Fauquet CM, et al. 2003. Revision of taxonomic criteria for species demarcation in the family Geminiviridae, and an updated list of begomovirus species. *Arch. Virol.* 148:405–421.
- Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (ed). 2005. *Virus taxonomy: eighth report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, Amsterdam, The Netherlands.
- Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351–360.
- González JM, Gomez-Puertas P, Cavanagh D, Gorbalenya AE, Enjuanes L. 2003. A comparative sequence analysis to revise the current taxonomy of the family Coronaviridae. *Arch. Virol.* 148:2207–2235.
- Gorbalenya AE, Lauber C. 2010. Origin and evolution of the Picornaviridae proteome, p 253–270. *In* Ehrenfeld E, Domingo E, Roos RP (ed), *The picornaviruses*. ASM Press, Washington, DC.
- Gorbalenya AE, et al. 2010. Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antiviral Res.* 87:95–110.
- Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis - detection of distantly related proteins. *Proc. Natl. Acad. Sci. U. S. A.* 84:4355–4358.
- Hebert PDN, Ratnasingham S, Dewaard JR. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. London B Biol. Sci.* 270:S96–S99.
- Hicks AL, Duffy S. 2011. Genus-specific substitution rate variability among picornaviruses. *J. Virol.* 85:7942–7947.
- Jiang P, et al. 2007. Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. *Proc. Natl. Acad. Sci. U. S. A.* 104:9457–9462.
- King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed). 2012. *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, Amsterdam, The Netherlands.
- Knowles NJ, et al. 2012. Family *Picornaviridae*, p 855–880. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), *Virus taxonomy: ninth report of the International Committee for the Taxonomy of Viruses*. Elsevier Academic Press, Amsterdam, The Netherlands.
- Knowles NJ, Hovi T, King AMQ, Stanway G. 2010. Overview of taxonomy, p 19–32. *In* Ehrenfeld E, Domingo E, Roos RP (ed), *The picornaviruses*. ASM Press, Washington, DC.
- Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23:1891–1901.
- Lauber C, Gorbalenya AE. 2012. Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses. *J. Virol.* 86:3905–3915.
- Lewis-Rogers N, Bendall ML, Crandall KA. 2009. Phylogenetic relationships and molecular adaptation dynamics of human rhinoviruses. *Mol. Biol. Evol.* 26:969–981.
- Lewis-Rogers N, Crandall KA. 2010. Evolution of Picornaviridae: an examination of phylogenetic relationships and cophylogeny. *Mol. Phylogenet. Evol.* 54:995–1005.
- Lewis-Rogers N, McClellan DA, Crandall KA. 2008. The evolution of foot-and-mouth disease virus: impacts of recombination and selection. *Infect. Genet. Evol.* 8:786–798.
- Li LL, et al. 2009. A novel picornavirus associated with gastroenteritis. *J. Virol.* 83:12002–12006.
- Maes P, et al. 2009. A proposal for new criteria for the classification of hantaviruses, based on S and M segment protein sequences. *Infect. Genet. Evol.* 9:813–820.
- Martin DP, et al. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.
- Martinez-Salas E, Ryan MD. 2010. Translation and protein processing, p 141–161. *In* Ehrenfeld E, Domingo E, Roos RP (ed), *The picornaviruses*. ASM Press, Washington, DC.
- Marvil P, et al. 1999. Avian encephalomyelitis virus is a picornavirus and is most closely related to hepatitis A virus. *J. Gen. Virol.* 80:653–662.
- Matthijnsens J, et al. 2008. Full genome-based classification of rotaviruses reveals a common origin between human Wa-like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. *J. Virol.* 82:3204–3219.
- Oberste MS, Maher K, Kilpatrick DR, Pallansch MA. 1999. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. *J. Virol.* 73:1941–1948.
- Palmenberg A, Neubauer D, Skern T. 2010. Genome organization and encoded proteins, p 3–17. *In* Ehrenfeld E, Domingo E, Roos RP (ed), *The picornaviruses*. ASM Press, Washington, DC.
- Perl Foundation. 2011. The Perl programming language. <http://www.perl.org>.

52. R Development Core Team. 2011. R: a language and environment for statistical computing. <http://www.R-project.org>.
53. Riquet FB, et al. 2008. Impact of exogenous sequences on the characteristics of an epidemic type 2 recombinant vaccine-derived poliovirus. *J. Virol.* 82:8927–8932.
54. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J. Virol.* 84:9733–9748.
55. Santti J, Hyypia T, Kinnunen L, Salminen M. 1999. Evidence of recombination among enteroviruses. *J. Virol.* 73:8741–8749.
56. Sayers EW, et al. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 38:D5–D16.
57. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
58. Schuffenecker I, Ando T, Thouvenot D, Lina B, Aymard M. 2001. Genetic classification of “Sapporo-like viruses.” *Arch. Virol.* 146:2115–2132.
59. Semler BL, Wimmer E (ed). 2002. Molecular biology of picornaviruses. ASM Press, Washington, DC.
60. Shukla DD, Ward CW. 1988. Amino-acid sequence homology of coat proteins as a basis for identification and classification of the potyvirus group. *J. Gen. Virol.* 69:2703–2710.
61. Sidorov IA, Samborskiy DV, Leontovich AM, Gorbalenya AE. 2012. HAYGENS, Homology-Annotation hYbrid retrieval of GENetic Sequences. <http://veb.lumc.nl/HAYGENS/index.cgi>.
62. Simmonds P. 2006. Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses. *J. Virol.* 80:11124–11140.
63. Simmonds P, Welch J. 2006. Frequency and dynamics of recombination within different species of human enteroviruses. *J. Virol.* 80:483–493.
64. Stanway G, et al. 2005. Family *Picornaviridae*, p 757–778. In Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (ed), *Virus taxonomy: eighth report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, Amsterdam, The Netherlands.
65. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56:564–577.
66. Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
67. Ward CD, Flanagan JB. 1992. Determination of the poliovirus RNA-polymerase error frequency at 8 sites in the viral genome. *J. Virol.* 66:3784–3793.
68. Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–699.
69. Wimmer E, Paul A. 2010. Making of a picornavirus genome, p 33–55. In Ehrenfeld E, Domingo E, Roos RP (ed), *The picornaviruses*. ASM Press, Washington, DC.
70. Wittkop T, et al. 2010. Partitioning biological data with transitivity clustering. *Nat. Methods* 7:419–420.
71. Zheng DP, et al. 2006. Norovirus classification and proposed strain nomenclature. *Virology* 346:312–323.
72. Zoll J, Galama JMD, van Kuppeveld FJM. 2009. Identification of potential recombination breakpoints in human parechoviruses. *J. Virol.* 83:3379–3383.