On the Evolution of Genetic Diversity in RNA Virus Species

Uncovering barriers to genetic divergence and gene length in picorna- and nidoviruses

Chris Lauber

front cover: ‘Out-of-the-box’. Genome sizes of ssRNA+ virus species are shown along a logarithmic spiral which reaches from 2 to 32 kilobases. The size ranges of picorna- and nidoviruses (light-colored balls) are indicated through yellow and turquoise shading, respectively. Concept by A.E. Gorbalenya; developed and implemented by C. Lauber using *R*.

back cover: ‘Perspective’. The text content of each page of the eight chapters in this thesis is shown. Inspired by a project of Ben Fry at his web site; implemented by C. Lauber using *Processing*.

The research described in this thesis was carried out at the Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands, and was financially supported in part by the Netherlands Bioinformatics Centre (BioRange SP 2.3.3) and the European Union (FP7 IP Silver HEALTH-2010-260644).

Funding of the printing costs by the Department of Medical Microbiology and the Netherlands Bioinformatics Centre (NBIC) is gratefully acknowledged.

On the Evolution of Genetic Diversity in RNA Virus Species

Uncovering barriers to genetic divergence and gene length in picorna- and nidoviruses

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 30 oktober 2012
klokke 16.15 uur

door

Chris Lauber

geboren te Gera, Germany
in 1981

Promotiecommissie

Promotor:	Prof.dr. A.E. Gorbalenya
Overige leden:	Dr. J.J. Goeman
	Prof.dr. J. Heringa *Vrije Universiteit Amsterdam*
	Prof.dr. F. van Kuppeveld *Universiteit Utrecht*
	Prof.dr. E.J. Snijder
	Prof.dr. W.J.M. Spaan

“Nothing in biology makes sense except in the light of evolution.”

Theodosius G. Dobzhansky, 1973

Table of contents

List of abbreviations

A	adenine
aa	amino acid(s)
C	cytosine
CO_2	carbon dioxide
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
G	guanine
HMM	Hidden Markov Model
ICTV	International Committee on Taxonomy of Viruses
kb	kilobases
ML	maximum likelihood
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NTP	nucleoside triphosphate
nt	nucleotide(s)
ORF	open reading frame
RNA	ribonucleic acid
ssDNA	single-stranded DNA
ssRNA+	single-stranded positive-sense RNA
ssRNA-	single-stranded negative-sense RNA
T	thymine
U	uracil
UTR	untranslated region

Samenvatting

Dit proefschrift combineert het gebruik van standaard bioinformatica analyses met de ontwikkeling van nieuwe computationele technieken om de evolutie en genetische diversiteit van picornavirussen en nidovirussen te bestuderen. Het integreert twee onderzoekslijnen - genetica gebaseerde virus classificatie en evolutionaire dynamiek van gen lengte - en richt zich op de onthulling van overeenkomsten in de biologie van deze en andere RNA virussen en op de ondersteuning van toegepast onderzoek in de virologie.

Hoofdstuk 1 introduceert basiskennis over RNA virus diversiteit, virus classificatie en standaard bioinformatica technieken die worden gebruikt in dit proefschrift. In **hoofdstuk 2** wordt een nieuwe kwantitatieve methode voor de hiërarchische classificatie van virussen door paarsgewijze genetische divergentie, genaamd *DEmARC*, beschreven en rigoureus geëvalueerd in een basisstudie van picornavirussen. DEmARC wordt toegepast in elke van de andere hoofdstukken, ofwel door de belangrijkste resultaten te produceren (hoofdstuk 3 en 4) of voor het maken van datasets die representatief zijn voor een veel grotere groep van virussen (hoofdstuk 5, 6 en 7). In **Hoofdstuk 3** blijkt DEmARC het opvallend eens te zijn met de officiële ICTV taxonomie van picornavirussen die wordt geproduceerd door gezamenlijke inspanningen van deskundige virologen en die regelmatig verfijnd wordt met behulp van verschillende kenmerken van virussen. Een paar opmerkelijke biologische verschillen en afwijkingen van de twee onderliggende classificatie benaderingen worden uiteengezet. In **hoofdstuk 4** wordt DEmARC voor het eerst gebruikt in een uitgebreide-familie-analyse om te helpen bij de indeling van de eerste nidovirussen die geïsoleerd zijn uit insecten (zie ook hoofdstuk 5). Deze virussen, genaamd *mesonivirussen*, bleken een aparte soort en vormen een nieuwe nidovirus familie. **Hoofdstuk 5** rapporteert over de ontdekking van een mesonivirus en haar onderscheidende genetische eigenschappen, met speciale nadruk op de genetische divergentie en de ongewone grootte van het genoom. Deze resultaten maakten de weg vrij om een model formeel te introduceren dat replicatie nauwkeurigheid, grootte van het genoom en genetische complexiteit verenigt; een grote beperkingen in de evolutie van RNA virussen. In **hoofdstuk 6** wordt dit model geconfronteerd met nidovirussen die genomen van extreme grootte geëvolueerd hebben. De gerapporteerde bevindingen tonen aan dat genomische architectuur een kritische factor vormt in de expansie van nidovirus genomen en mogelijk de waargenomen genoom limiet voor de hedendaagse RNA virussen kan bepalen. **Hoofdstuk 7** beoordeelt onze huidige kennis van de evolutie van de picornavirus eiwitten. Zoals voor nidovirussen, de afwijking van genetische elementen wordt geanalyseerd in de context van de sequentie en grootte. Uit deze bevindingen blijkt dat er een negatieve correlatie van sequentie behoud en de grootte variatie van picornavirus eiwitten bestaat. **Hoofdstuk 8** sluit dit proefschrift af door het bespreken van zijn bevindingen en implicaties voor toekomstig onderzoek.

Outline of this thesis

This thesis combines the use of standard bioinformatics analyses with the development of new computational techniques to study the evolution and genetic diversity of picornaviruses and nidoviruses. It integrates two lines of research – genetics-based virus classification and evolutionary dynamics of gene length – and aims at unveiling commonalities in the biology of these and other RNA viruses as well as assisting applied research in virology.

Chapter 1 introduces basic knowledge about RNA virus diversity, current virus classification efforts, and standard bioinformatics techniques utilized in this thesis. In **chapter 2** a novel, quantitative framework for hierarchical classification of viruses by pairwise genetic divergence, named *DEmARC*, is described and evaluated rigorously with respect to various key parameters in a proof-of-principle study of picornaviruses. DEmARC is employed in any of the other research chapters either to produce main results (chapter 3 and 4) or for the preparation of datasets that are representative for a much larger bunch of viruses (chapter 5, 6 and 7). In **chapter 3** DEmARC is shown to agree strikingly on the official ICTV taxonomy of picornaviruses which is produced by cooperative efforts of expert virologists who refine and update it periodically using various virus characteristics. A few biologically notable discrepancies as well as differences of the two underlying classification approaches are outlined. In **chapter 4** DEmARC is used, for the first time, in a multi-family analysis to assist the classification of the first nidoviruses isolated from insects (see also chapter 5). These viruses, named *mesoniviruses*, were found to form a single species prototyping a novel nidovirus family. **Chapter 5** reports on the discovery of one mesonivirus and its distinctive genetic properties in relation to other nidoviruses, with special emphasis placed on genetic divergence and the genome size of this virus which is unique among nidoviruses. The reported findings paved the way to formally introduce a model that unites replication fidelity, genome size and genetic complexity, major constraints in the evolution of RNA viruses. In **chapter 6** this model is contrasted with nidoviruses which evolved genomes of extreme size, thereby exceeding limits on the abovementioned constraints observed by other RNA viruses. The reported findings reveal that genomic architecture constitutes a critical factor in the expansion of nidovirus genomes and may determine the genome size limit observed for contemporary RNA viruses. **Chapter 7** reviews our current knowledge of the evolution of the picornavirus proteome. Like for nidoviruses, the divergence of genetic elements is analyzed in the context of sequence and size. The findings reveal a negative correlation of sequence conservation and size variation of picornavirus proteins. Finally, **chapter 8** concludes this thesis by discussing its findings and outlining implications for future research.

CHAPTER 1

General Introduction

RNA virus diversity in a nutshell

One of the most fundamental characteristics of RNA viruses is genetic variation. It originates from different sources which include (i) mutation (misincorporation of nucleotides during genome copying), (ii) duplication of genome regions (e.g. genes) followed by subsequent diversification, (iii) acquisition of foreign genetic material, (iv) loss of genetic material, and (v) overprinting (opening of an alternative, possibly overlapping reading frame)[34,252]. The underlying molecular processes act on different scales and may have profoundly different impacts on the virus affected.

At the lowermost level, genetic variation in RNA viruses is observed within single-host infections where a cloud of potentially beneficial mutations is formed and maintained. The cloud sequences can be thought of representing a fitness landscape – sequences with slightly lower or higher fitness – which allows the virus to quickly respond to environmental changes during infection. What is obtained by genome sequencing is a so-called *master sequence* which basically represents a consensus over the cloud. Interestingly, when infecting a new individual with the same virus, essentially the same cloud is formed which suggests that the sequences, linked by mutational coupling, act as one entity which is targeted as a whole by evolutionary selection[43]. This prompted researchers to refer to it as *Quasispecies*[110,131,132,471]. However, others are reluctant to accept the applicability of this concept to RNA viruses[219,237]. Regardless of this debate, it is apparent that genetic variation naturally increases on higher levels, e.g. when distinct viruses diverge gradually during evolution, eventually erasing any detectable homology. The reason for this high sequence variability of RNA viruses is the extreme error rate of their RNA-dependent polymerases, which introduce roughly one mutation per 10,000 nt copied[117,121]. In contrast to their cellular counterparts with an error rate that is orders of magnitude lower, polymerases of RNA viruses lack a proofreading-repair functionality (with one possible exception, see below).

Besides mutation, the second major source of genetic variation in RNA viruses is recombination. Recombination is thought to be guided by the level of local sequence similarity of the participant nucleic acid molecules[492] and, hence, is predominantly observed between closely related viruses. As a result, a recombinant genome in which homologous parts have been exchanged between the parents is produced (homologous recombination), potentially causing several substitutions in one go with respect to either parent. A prerequisite for homologous recombination to take place is co-infection of the same cell. It is worth mentioning that, while representing a source of innovation on a small scale, homologous recombination limits sequence divergence on a large scale (evolutionary time scale) by both the propagation of beneficial substitutions throughout a population of closely related viruses and the removal of deleterious mutations[305]. Moreover, recombination may also happen with remotely related viruses or the host transcriptome, and as a result new genes or other functional elements are integrated into the recipient viral genome (non-homologous recombination). In contrast to the mechanism of recombination in the host

genome which works through breakage and rejoining of DNA strands[80,251], it is generally accepted that that of most RNA viruses is characterized by template switching of the viral polymerase during replication[6,260,329].

With increasing genetic divergence, RNA virus diversity is also observed at the level of the proteome. A hallmark of RNA viruses and viruses in general is that they do not encode ribosomes and are thus condemned to parasitize host cells for protein synthesis. However, they express various other types of proteins. The only recurrent theme here is that all *bona fide* RNA viruses encode at least one virion protein as well as a polymerase, for genome dissemination and replication, respectively. The polymerase can be of different type depending on the mechanism of mRNA production (see next section). Its core component, the palm subdomain, which is crucial for catalysis, is structurally conserved across RNA viruses[184,367]. Many but not all RNA viruses express one or several proteinases which are utilized either for cleavage of viral polyproteins or to interact with the host environment[173]. RNA virus proteinases are grouped, on the basis of sequence similarity, into the chymotrypsin-related and papain-related superfamilies and proteases of unknown type. The chymotrypsin-related superfamily is further divided into the 3C-like family (named after the picornavirus proteinase encoded in the 3C genome region) and the CP-like family (named after the alphavirus capsid protein)[175]. Another type of enzyme frequently found in the RNA virus proteome is a helicase, characterized by containing a NTP-binding sequence pattern. It provides an activity – the unwinding of dsRNA or base-paired ssRNA – that is crucial to many genetic processes like replication, expression or recombination[177]. RNA virus helicases are grouped, on the basis of sequence similarity, into three superfamilies, two of which also include cellular proteins[175,177]. Besides these widespread enzymes, various other types of proteins are employed by RNA viruses including, for instance, ribonucleases, methyltransferases, phosphatases or membrane-associated proteins[73,96,128,150,174,376]. Notably, many RNA virus-encoded proteins have so far not been characterized functionally, partially due to the lack of any apparent similarity with cellular counterparts. Owing to the limited coding capacity of RNA viruses most of their proteins are multifunctional, but, at the same time, there is a certain division of labor between the enzymes[7].

More fundamentally, RNA virus diversity is observed at the level of the genome. RNA viruses vary in employing one of three possible genome types – double stranded (dsRNA viruses), single-stranded negative sense (ssRNA- viruses) or single-stranded positive sense (ssRNA+ viruses and RNA retroviruses). The latter represents by far the most abundant genome type and outnumbers by around fivefold each of the other two[25]. What's more, RNA virus genomes show a considerable variability in size[174]. Despite being limited to relatively small sizes compared to cellular organisms and most DNA viruses, RNA virus genomes still range from approximately 2 to 32 kb with an average of about 10 kb. Larger genomes are supposed to result in a so-called *error catastrophe* caused by too many deleterious mutations that would accumulate during the error-prone copying of RNA genomes[216]. Yet, the diversity of RNA genome sizes is striking and it was proposed that the

acquisition of specific enzymes that refined the rudimentary viral replication machinery empowered some RNA viruses to undergo major evolutionary transitions accompanied with an enlargement of the genome beyond the size of the acquired gene[174].

Last but not least, there is variation among RNA viruses from an economic and medical perspective. Most RNA viruses have rather mild effects on their hosts, which makes sense from an evolutionary perspective because the virus should avoid compromising its means of existence. However, there are also numerous viral pathogens that present serious threats to health care and economy. Examples are influenza A virus, measles virus, rinderpest virus or ebolavirus (ssRNA-), as well as rotaviruses and bluetongue virus (dsRNA). The most abundant viruses with a significant impact on society, however, are those with ssRNA+ genomes and include poliovirus, foot-and-mouth disease virus, tobacco mosaic virus, hepatitis C virus, yellow fever virus, dengue fever virus, noroviruses, or severe acute respiratory syndrome (SARS) coronavirus, to name only few. The focus of this thesis lies mainly, but not exclusively, on picorna- and nidoviruses.

Picornaviruses are among the most extensively studied and best characterized viruses. The first human virus (poliovirus) and the first animal virus (foot-and-mouth disease virus) to be discovered are picornaviruses. They employ small, non-enveloped virions and a ssRNA+ genome of around 6.5-9 kb length (Fig. 1A). The genome directly acts as an mRNA to encode (with few exceptions) a single polyprotein that is cleaved into a dozen or so mature proteins[480]. Among these viral proteins are three to four virion proteins, a superfamily 3 helicase with putative activity (denoted 2C in Fig. 1A), a chymotrypsin-related proteinase (3C), a RNA-dependent RNA polymerase (RdRp; 3D), and several other proteins that are deployed mainly to interact with the host environment. Picornaviruses can cause various, mostly acute diseases in animals and humans including poliomyelitis, foot-and-mouth disease, common cold, gastroenteritis, hepatitis, meningitis, myocarditis and uveitis[130]. Picornavirus studies contributed crucially to the general understanding of various aspects of (viral and cellular) molecular biology and virus-host interactions[7].

Also nidoviruses are known to infect only animals and to employ a ssRNA+ genome. They often cause fatal diseases like SARS in humans, feline infectious peritonitis in cats or infectious bronchitis of chickens and yellow head disease of prawns in livestock. Nidoviruses differ tremendously from picornaviruses in many aspects. The genome is organized in multiple ORFs (Fig. 1B), of which the biggest two are expressed directly from the genomic RNA[58] involving a ribosomal frameshifting event[56]. The other, smaller ORFs, whose number vary between three[429] to twelve[432], are expressed from subgenomic mRNAs that are synthesized by discontinuous extension during subgenome-length minus-strand synthesis[408]. Only three enzymes are common to all nidoviruses which are a chymotrypsin-related 3C-like proteinase (3CLpro), a superfamily 1 helicase (HEL1), and a RdRp. The nidoviral 3CLpro and RdRp show detectable sequence similarity to their homologs in picornaviruses[175,184,433].

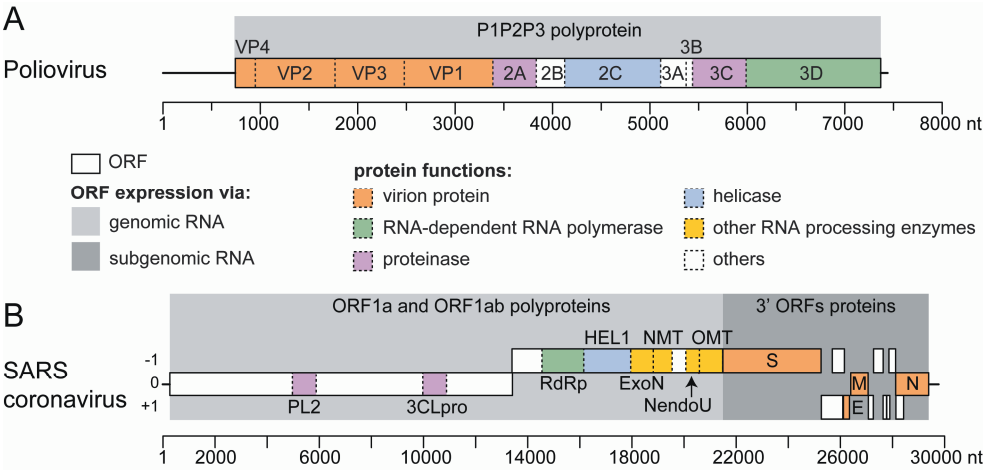


Figure 1. Genomic organization of picorna- and nidoviruses. The ssRNA+ genomes of poliovirus (A) and SARS coronavirus (B) are shown in 5'-to-3' direction and drawn to different scales. ORFs are depicted as bold rectangles and 5' and 3' untranslated regions as horizontal lines. The ORF expression mechanisms for (poly)protein production are indicated by grey background shading, and a selected set of protein functions is highlighted by coloring. An equal color does not necessarily imply sequence homology. (A) The genome of most picornaviruses expresses a single polyprotein that is cleaved into mature products by one or several viral proteinases. Picornaviruses other than poliovirus may differ in the polyprotein regions 2A and 3B and may additionally encode one or several leader proteins upstream of VP4. (B) The nidovirus genome encodes multiple, potentially overlapping ORFs in all three reading frames (-1, 0, +1). Using the 5'-proximal ORFs 1a and 1b two large polyproteins, pp1a and pp1ab, are expressed. The production of the latter involves a -1 ribosomal frameshifting event. The 3'-proximal ORFs are expressed via multiple subgenomic mRNAs to produce virion and accessory proteins. Only three enzymes - chymotrypsin-related 3C-like proteinase (3CLpro), RNA-dependent RNA polymerase (RdRp) and superfamily 1 helicase (HEL1) - are conserved across nidoviruses. Other highlighted proteins include: papain-related proteinase (PL2), exoribonuclease (ExoN), N7-methyltransferase (NMT), uridylate-specific endoribonuclease (NendoU), 2'O-methyltransferase (OMT), spike protein (S), envelope protein (E), matrix protein (M), and nucleocapsid protein (N).

Other proteins encoded by some nidoviruses have diverse functionalities including endo- and exoribonuclease, methyltransferase or phosphatase activities[174] that are rarely or never observed outside this virus group. The extraordinary genetic complexity of nidoviruses is reflected in their large genomes whose sizes are well above that of the average RNA virus genome. The smallest nidovirus genome is approximately 12.5 kb in size whereas the largest ones reach almost 32 kb which, in fact, makes them the largest RNA genomes known to date. Interestingly, nidovirus genomes of sizes above 20 kb are uniquely associated with the expression of a special enzyme, a 3'-to-5' exoribonuclease of the DEDD superfamily. This enzyme was suggested to improve the fidelity of RNA replication[432], a role which is supported by three independent lines of evidence. First, it is distantly related in sequence to a cellular proofreading enzyme. Second, the gene encoding the exoribonuclease is genetically segregated with genes that encode key enzymes of the nidovirus replication complex. And third, its functional activity was shown for mouse hepatitis

virus[124] and SARS coronavirus[123]. The extreme genome size and complex proteome composition make nidoviruses an exciting model to study the relation of genome size and mutation rate in RNA viruses[174] and, potentially, in biology in general.

Whys and wherefores of virus classification

Having in mind that tremendous diversity of RNA viruses briefly outlined in the previous section (as well as that of their DNA cousins), it comes not as a surprise that scientists aimed to structure the *Virosphere* ever since these infectious agents came to the fore in science at the end of the 19th century[30,233,301]. In order to do so, a classification has to be devised which, as a basic principle, groups together similar objects in a biologically meaningful way. A classification relies on one or several distinctive characteristics which are often referred to as *demarcation criteria*.

One such criterion is the mode of viral mRNA production. It is related to the viral genome type and defines a coarse-grained classification that consists of seven virus classes[23]. Three of them are formed by viruses with DNA genomes that produce mRNA either by direct transcription of the genome (dsDNA viruses), by involving the formation of a double-stranded intermediate (ssDNA), or by first filling their gapped genomes using reverse transcription (dsDNA with RNA intermediate). The remaining four virus classes employ RNA genomes that serve as a template for mRNA synthesis (dsRNA and ssRNA-), directly act as mRNA (ssRNA+), or require a DNA intermediate which is integrated into the host genome to subsequently undergo the cellular transcription process (retroviruses). The ssRNA+ genome type is by far the most abundant and outnumbers around fivefold each of the other types[374]. Since this classification scheme ignores most variety observed for viruses, it is suitable only for very general purposes.

The most widely used form of virus classification, as with cellular organisms, is *taxonomy*. Virus taxonomy presents a hierarchical system which comprises the ranks, from top to bottom, order, family, subfamily, genus, and species[257]. The two most crucial layers here are the family and species ranks due to different reasons. A family is used for grouping together viruses that share some rather general properties which uniquely discriminate them from other families. For each virus family there is a group of specialized virologists who propose, update or revise taxonomic entities largely independently for that particular family. These so-called *virus study groups* operate under the direction of the *International Committee on Taxonomy of Viruses* (ICTV) which, in turn, is administered by the Virology Division of the International Union of Microbiological Societies. The ICTV is the (only) official body with the legitimation to approve taxonomic assignments of viruses. According to the latest taxonomy release in 2012, 94 virus families are currently recognized by the ICTV[257]. Virus species, on the other hand, are the basic taxonomic entities[373] and imply highly similar virus phenotypes. They are defined as “a polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche”[257]. With other words, for virus

species demarcation it is taken into consideration a broad range of characteristics, including genetic and phenotypic properties, each of which is shared by some but not necessarily all members of the species – a form of consensus decision. Currently, 2475 different virus species are recognized by the ICTV[257]. The ICTV offers an extremely powerful and flexible framework for virus classification which can accommodate any virus diversity. However, this comes with the costs of a substantial amount of scientists' labor time and a relatively lengthy decision process. Names of ICTV-defined taxa are written in italics to discriminate them from virus strains or isolates.

The expert-based ICTV classification framework might approach its logistic limitations in the near future due to the rapid increase of newly described viral genome sequences[358] (Fig.2). Consequently, ICTV study groups increasingly explore the usability of genetic data for decision making in virus taxonomy[264]. Moreover, several research groups proposed a purely genetic-based classification for a virus family or genus (see Introduction in chapter 2). They thereby exploited the fact that genome sequencing greatly outpaces all other types of virus characterization and produces high-quality data which is most suitable for quantitative analysis. A recurrent theme in these analyses is the utilization of a so-called *pairwise distance distribution* formed by genetic distances between all pairs of viruses under consideration. Often, the percentage of sequence identity is used as a measure of pairwise distance. With this sequence-based approach a classification is derived by partitioning the pairwise distance distribution into intra-rank (e.g. intra-species) and inter-rank regions. For example, all distances below a certain threshold are attributed to viruses from the same species and those above the threshold to virus pairs from different species. In doing so, it is assumed that there exist common factors that result in the separation of taxonomic ranks on a genetic level. The crucial challenge which remains is to define the number and position of distance thresholds. Preferably, this should be done in an objective manner which, however, is a goal not aimed at by current approaches.

RNA virus classification makes a difference

RNA virus classification efforts brought some insightful findings with great impact on virology. Perhaps one of the most illustrative examples is the picornavirus species *Human enterovirus C*[263] which includes the major human pathogen poliovirus (PV) as well as eleven serotypes of the rather benign C-cluster coxsackie A viruses (CCAVs). Initially, PV and CCAVs were classified as two separate species[436]. Largely due to sequence-based phylogenetic analyses, it was found that the genetic similarity of PV and CCAVs is large enough to form a single species[238,263]. In line with that are recombinants between PV and CCAVs that have been identified in the field and are mostly vaccine-dervied[21,59,390]. These findings have great implications for the Global Polio Eradication Initiative[238], a campaign to eradicate poliovirus from our planet which was initiated in 1988 but still did not succeed globally[112], and it is unclear if it ever will[331].

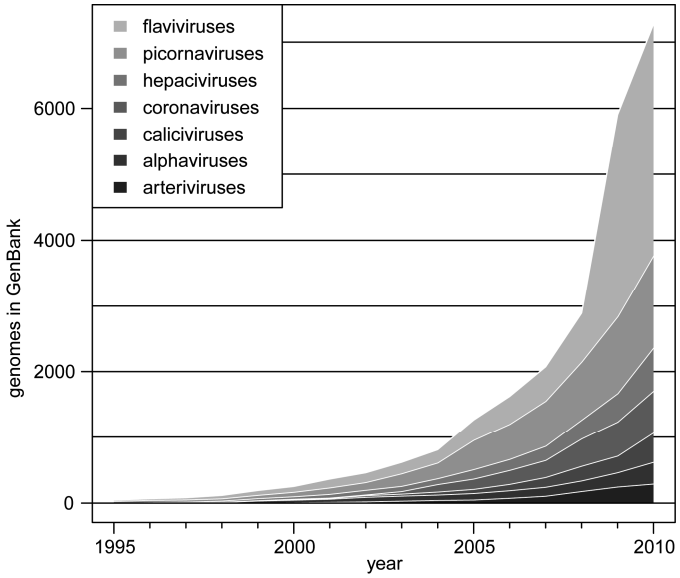


Figure 2. Non-linear increase in the number of ssRNA+ viruses with sequenced genomes. The accumulation of complete genome sequences in the VRL and PAT divisions of GenBank during the last one and a half decade is shown for selected ssRNA+ virus groups (see legend). Genome sequences have been obtained and grouped using the HAYGENS tool[423].

Another example for the potential value a virus classification can have is the species *Theilovirus*[263] that was initially thought to infect only rodents[206,343,449]. Later, phylogenetically closely related viruses have been isolated from humans and it was shown that the rodent and human lineages are genetically similar enough to be members of the same species, basing on the comparison with other picornavirus species[161,299]. This indicates that a virus species is not bound to infect a single host species, a finding supported by phylogenetic analyses of other virus families[262].

Furthermore, from a more technical perspective, a virus classification provides means for the reduction of a dataset in size in order to perform certain resource-consuming analyses. For phylogeny reconstruction (see next section), for instance, it is often necessary to chose representatives from a much larger set of viral sequences. Depending on the evolutionary scale of the analysis, virus species may present a meaningful and objective choice. In the case of picornaviruses, it allows for decreasing the dataset of available complete genomes by more than one order of magnitude from hundreds of sequences to a few dozen species. As a second effect, such a virus selection per species partially corrects for the inherent sampling bias of a dataset which can be enormous and is caused by the difference in medical or economical importance of the viruses. *Foot-and-mouth disease virus* (FMDV)[109,265] for example, one of the most economically important and best studied RNA viruses, shows the highest sampling among all picornavirus species with a few hundred

complete genomes available, whereas other species of that family, like *Seneca Valley virus* (SVV)[197], are represented only by a single sequence.

Bioinformatics meets virology

All bioinformatics analyses utilized in this thesis depend in one way or another on sequencing data which is accessible from public databases like GenBank/RefSeq[36]. In virology this type of information presents an invaluable and virtually infinite resource nowadays (Fig.2), however, one should keep in mind that it comes with certain challenges due to the nature of viruses and their elevated rate of introducing sequence innovations. Various bioinformatics techniques are available that can withstand or at least account for theses challenges, some of which have been utilized in one or several of the following chapters. They are briefly discussed below.

A commonly used tool to compare related biological sequences is a multiple sequence alignment. It is typically built by maximizing the similarity among the sequences, which involves the introduction of gaps at sequence positions that are not conserved. These gap positions predict insertion/deletion (indel) events that have happened during evolution, whereas non-gap positions infer common ancestry (homology) of the respective sequence residues. For an alignment of highly diverged sequences, like that of a typical evolutionary study of RNA viruses, it is common to partition it into conserved regions, so-called *blocks*, and poorly conserved parts[18,65]. The latter are prone to contain alignment artifacts due to an elevated substitution and/or indel rate in these regions, and, hence, may need to be discarded from subsequent analyses. Multiple sequence alignments build the foundation for many other types of bioinformatics analyses. For example, they can be converted to profiles, which are statistical models and capture for each alignment column the degree of conservation and the likelihood to observe a certain residue or gap. One type of profiles are profile HMMs[125,273,434] which operate inside a probabilistic framework and are particularly suitable for the detection of remote sequence homology. A profile HMM can be compared to other HMMs or used to search for motifs in a single sequence.

Another approach to detect or support remote homology, especially when sequence divergence erased most similarity, is structure predictions, which can be applied to both RNA and protein sequences. In the case of RNA, the folding that results in the minimal free energy is considered biologically most meaningful and desirable. A folding is defined by basepairing interactions (stems) and unpaired regions (loops), and may include pseudoknots (interaction of two or more stem-loop structures). Pseudoknots and other structural elements are often found in UTRs of RNA viruses[57,389]. For protein sequences, three basic secondary structure elements are distinguished – alpha helices, beta sheets, and loops. During prediction, each sequence residue is assigned to one of these three states, thereby taking into account physiochemical properties of the amino acids, for

instances hydrophobicity. Protein secondary structure predication can be combined with profile searches (see above) to improve the sensitivity of the latter[434].

Assuming that sequence homology is still detectable, a widely used approach to infer the relatedness of a set of sequences is phylogeny reconstruction. Typically starting from a multiple alignment, it results in an evolutionary tree that represents order and degree of sequence divergence. The tree can be in unrooted or rooted form depending on the availability of additional information, often brought by an outgroup (a sister lineage related to the sequences of interest). Importantly, certain assumptions have to be met in order to obtain meaningful results from phylogenetic analyses. It includes common ancestry of the sequences under consideration, neutrality of the vast majority of molecular changes[256], and adequate approximation of the true substitution patterns by the evolutionary model used[275]. There are several frameworks for phylogeny reconstruction, one of which is distance methods. Here, for each pair of sequences an evolutionary distance is computed and the relationships among these values are used to gradually group the sequences starting with the most closely related pair. Another popular and speedy technique is Maximum Parsimony (MP), which takes a greedy approach by seeking to find the tree that would result in the least number of substitutions throughout the phylogeny[122,152,201]. Maximum Likelihood (ML) is one of the most powerful reconstruction techniques, as being a probabilistic framework[66,145]. Under a specific substitution model, ML identifies the tree that maximizes the likelihood to observe the given set of sequences. Importantly, the distances between sequences, so-called *branch lengths*, are a parameter of the method and are thus optimized together with the branching order, the *topology*. Finally, there is Bayesian methods which are related to ML[381]. They allow for the incorporation of external knowledge, so-called *priors*. Prior knowledge could be, for instance, a known substitution rate or the monophyly of a subset of sequences. Furthermore, Bayesian methods provide confidence intervals for every parameter estimated, which represents a major advantage over the other methods. ML and Bayesian frameworks are considered more sophisticated and hence were favored throughout this thesis.

As indicated in the FMDV-SVV example above (see previous section), sampling bias is a common problem in sequence-based bioinformatics. In certain situations it may distort or even obscure signals in the data. Fortunately, such effects can be circumvented or at least diminished by using sequence weights. As a consequence, the sequences contribute unequally to the results, with most unique sequences having the highest impact. There are several approaches how to calculate these weights[157,209,472].

One of many situations where sequence weights can be of high value is regression analysis. For instance, it could be of interest to determine if there is any relation between two parameters shared by all sequences of a dataset (a naive example would be sequence length compared to G+C content). Regression analysis is a general statistical technique to model the relationship between a dependent variable and one or more independent variables. Such a relation, if any exists, can be of linear or non-linear nature and is often

referred to as *correlation*. Once a correlation has been established, regression analysis can be used to make predictions for new data points (e.g. new sequences). It is important to note that correlation does not necessarily imply causation. For example, since the middle of the last century, both atmospheric CO_2 level and obesity level increased strongly. Though, one might not infer that atmospheric CO_2 causes obesity. Instead, these two parameters could be linked by an increase in car sales.

Scientific questions

The first part of this thesis is devoted to RNA virus classification in an evolutionary context. Specifically, it is asked whether viruses of a family can be classified objectively and accurately by basing solely on genome sequences, the footprint of evolution, and whether this approach can be extended to multi-family analyses. It is further explored whether such a classification can deliver novel insight into constraints on genetic diversity in RNA viruses and what benefit for applied research in virology it can provide.

The second part is devoted to the analysis of gene and genome size change during RNA virus evolution. It is asked whether there are factors other than the fidelity of replication that limit genetic size in RNA viruses, what principal proteins are involved in the control of genetic size, and whether size change is linked to the accumulation of mutations. Moreover, it is sought to unravel biological factors that drove the emergence of the largest known RNA genomes employed by nidoviruses.

CHAPTER 2

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

Chris Lauber
Alexander E. Gorbalenya

Journal of Virology (2012) 86:3890
(JVI **spotlight** feature)
(selected by **Faculty of 1000**)

Abstract

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[283], we analyze the substantial insight gained with the genetics-based classification approach by comparing it with the expert-based picornavirus taxonomy.

Introduction

Viruses form a large class of biological entities of extreme diversity[107]. 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)[143,257]. 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[36,410]. For a growing number of viruses, the genome sequence is the first and often the only information available (for a review, see references [88,97,127]). 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[257].

Another complementary approach that is steadily growing in popularity is so-called pairwise sequence comparisons (pasc)[422]. 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)[24]. 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[1,2,17,38,94,142,162,309,317,413,422,496].

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[293]. Since RNA viruses are known for the extremely high mutation rates of their polymerases[117,121,473], 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 "*D*iv*E*rsity p*A*rtitioning by hie*R*archical *C*lustering" and refers to the English word "demarcation." We extensively tested DEmARC on the proteome of the *Picornaviridae*[182], one of the most diverse and well-studied RNA virus families[130,415] with numerous species that has been developed by one of the most active SGs[264,265,435]. The picornavirus genome is a single-stranded positive-sense RNA (ssRNA+) with a single open reading frame that encodes a polyprotein[313,346] flanked by two untranslated regions, 5′-UTR and 3′-UTR[481]. 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[283], we analyze implications of the developed genetics-based classification for fundamental and applied research, through its comparison with virus phylogeny and taxonomy.

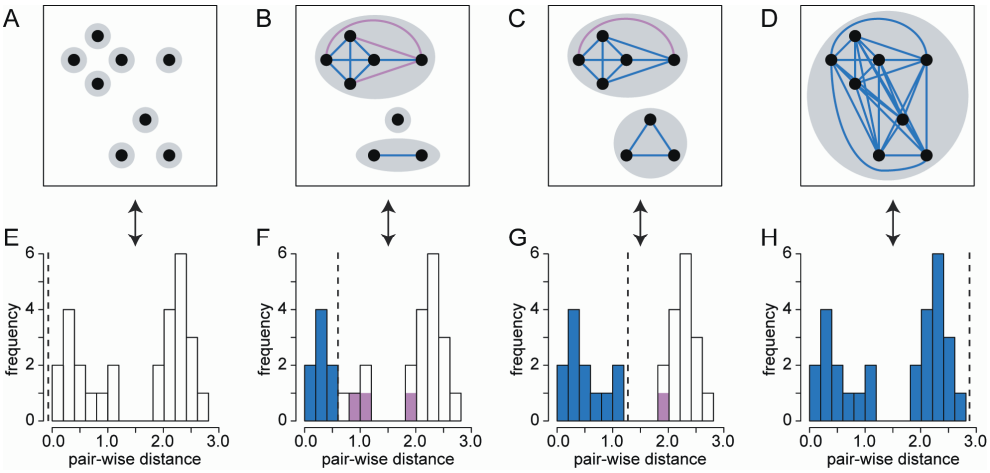


Figure 1. Grouping viruses based on thresholds in the distribution of pairwise genetic divergence. Shown is a fictitious example involving eight viruses that illustrates the relation between the selection of a threshold in the distribution of intervirus genetic divergence and the accompanied change in 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.

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[36] databases were downloaded using HAYGENS[423] into the Viralis platform[183]. A multiple-amino-acid alignment of the polyproteins was produced using the Muscle program version 3.52[126], and poorly conserved columns 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[313].

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 interspecies 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[18,65,443]. These blocks constitute regions of highest alignment quality/accuracy and account for ~63% alignment positions 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[126] or ClustalW version 2.0.12[452] program (E-Muscle and E-Clustal, respectively).

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	+	-	-	-	Picorna	1BCD, 2C, 3CD	2010	1234	1543
E-Capsid	+	-	-	-	Picorna	1BCD	2010	1234	1246
E-G1	+	+	-	-	Entero, sapelo	P1, 2BC, P3	2010	706	2322
E-G2	+	+	-	-	Avihepato	P1, P2, P3	2010	65	2255
E-G3	+	+	-	-	Hepato, tremo	P1, 2BC, P3	2010	58	2060
E-G4	+	+	-	-	Parecho	P1, 2A6BC, P3	2010	44	2302
E-G5	+	+	-	-	Kobu, sali*	P1, P2, P3	2010	12	2401
E-G6	+	+	-	-	Aphtho	L, P1, P2, 3AB2CD	2010	267	2480
E-G7	+	+	-	-	Cardio, seneca	P1, 2A4BC, P3	2010	39	2219
E-G8	+	+	-	-	Tescho	L, P1, P2, P3	2010	31	2242
E-G9	+	+	-	-	Cosa*	P1, P2, P3	2010	9	2154
E-G10	+	+	-	-	Avihepato, parecho, aquama*	P1, 2A4BC, 3AB2CD	2010	110	2312
E-G11	+	+	-	-	Aphtho, erbo, cardio, seneca, cosa*, tescho	P1, 2A4BC, 3AB2CD	2010	348	2748
E-2008	-	+	-	-	Picorna	1BCD, 2C, 3CD	2008	685	2374
E-2006	-	+	-	-	Picorna	1BCD, 2C, 3CD	2006	427	2280
E-2004	-	+	-	-	Picorna	1BCD, 2C, 3CD	2004	181	2269
E-Muscle	-	-	+	-	Picorna	1BCD, 2C, 3CD	2010	1234	2592
E-Clustal	-	-	+	-	Picorna	1BCD, 2C, 3CD	2010	1234	2269
E-PUD	-	-	-	+	Picorna	1BCD, 2C, 3CD	2010	1234	2446
E-PASC	+	-	+	+	Picorna	Complete genomes	2010	1234	-[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 1234 sequences and 2446 positions.

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

[d] P1, P2, P3 comprise capsid proteins (1A to 1D), non-structural part 1 (2A to 2C), non-structural part 2 (3A to 3D), respectively; for 2A and 3B designations see[182].

[e] Sampling date of data set according to Genbank annotation.

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

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[412]. The WAG amino acid substitution matrix[478] 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. Consequently 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.

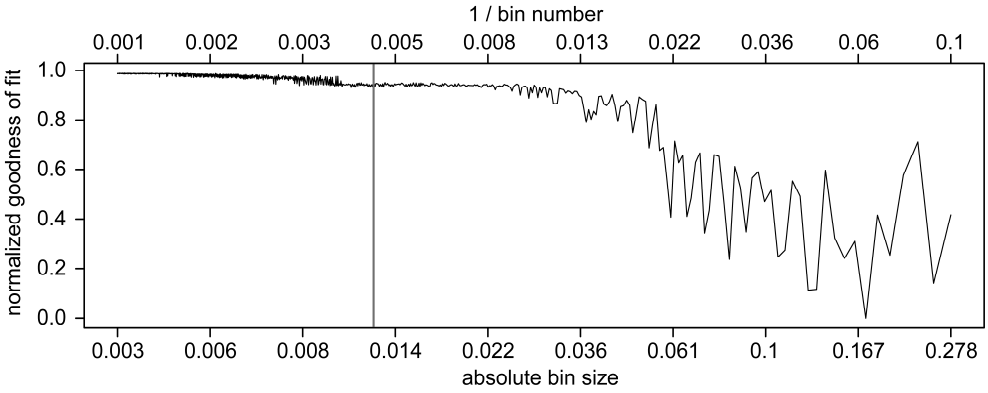


Figure 2. Optimal bin number for the picornavirus-wide pairwise distance distribution. Shown is the χ^2 goodness-of-fit measure for approximating the picornavirus-wide PED distribution with normal probability densities using different bin sizes. Ten to 1,000 bins were tested, and the measure was normalized 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, $\alpha = 0.01$.

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 [101] 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

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

with fk being the probability density function that approximates component *k* for (*k = 1,…,K − 1*)-determined peak components and the background component, component weights $w_1,\ldots,w_K$ (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:

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

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:

$$GOF = 1 - \frac{\chi^2}{b} \quad (3)$$

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

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

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:

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

for inferred clusters $c = 1,\ldots,C$, intragroup distance values d_c, and distance threshold value t. The CC is a simplification of the modification cost defined in reference [483], 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[359] and R[377] 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.

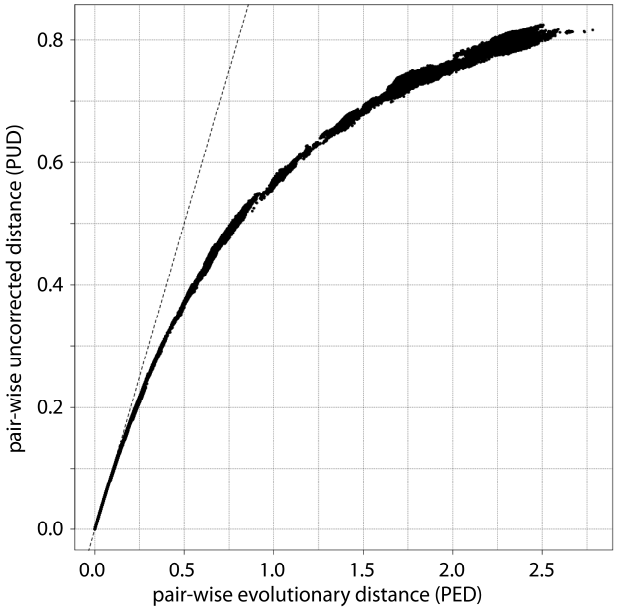


Figure 3. Corrected versus uncorrected picornavirus-wide pairwise distances. Plotted is corrected pairwise evolutionary distance (PED) versus pairwise uncorrected distance (PUD) for the M-2010 data set. For intermediate and large distances, a saturation of PUD values is observed, as they do not account 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.

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.

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

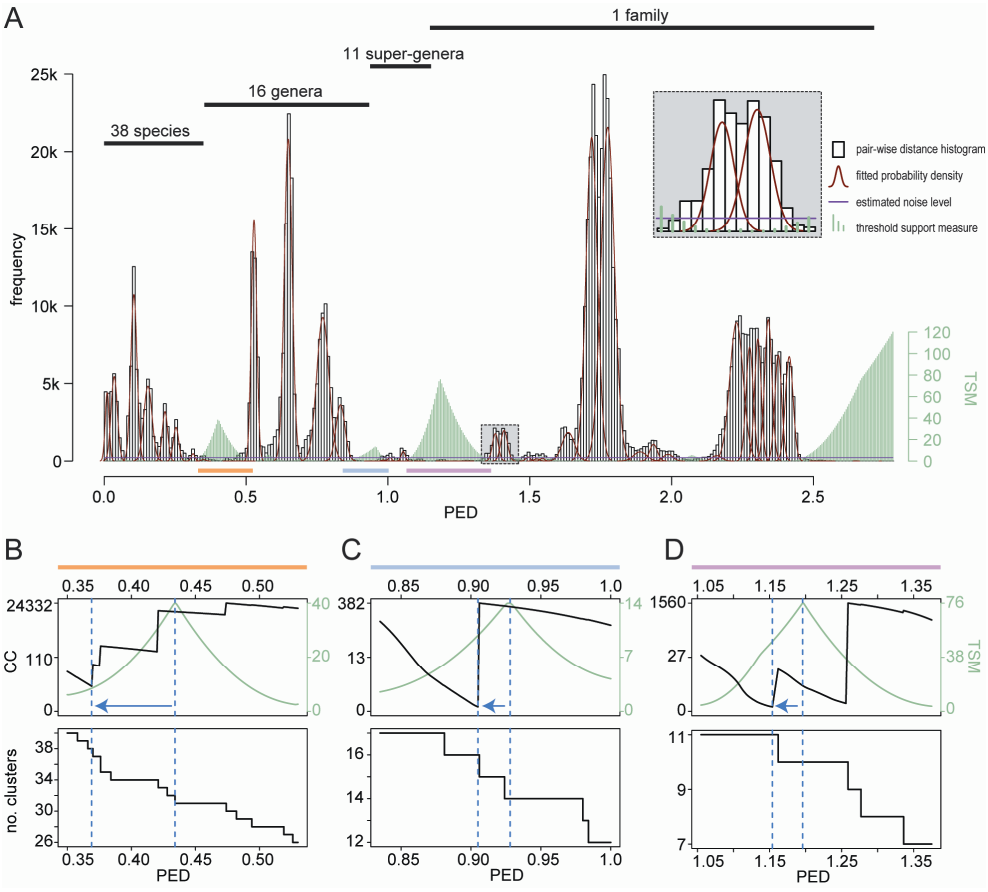


Figure 4. Picornavirus-wide pairwise distance distribution and distance thresholds for partitioning. (A) Frequency distribution of ~760,000 PED values is shown for the M-2010 data set. In a first stage (see inset), peaks in the distribution were approximated using a mixture of normal distributions (red 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.

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

	Species			Genus			Supergenus		
Data set[a]	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 data sets.

[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 data sets in respect to this virus subset.

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

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)[283].

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 ≤ CA ≤ 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 columns in M-2010 with the lowest conservation scores[18,65] 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 (supergenus), respectively.

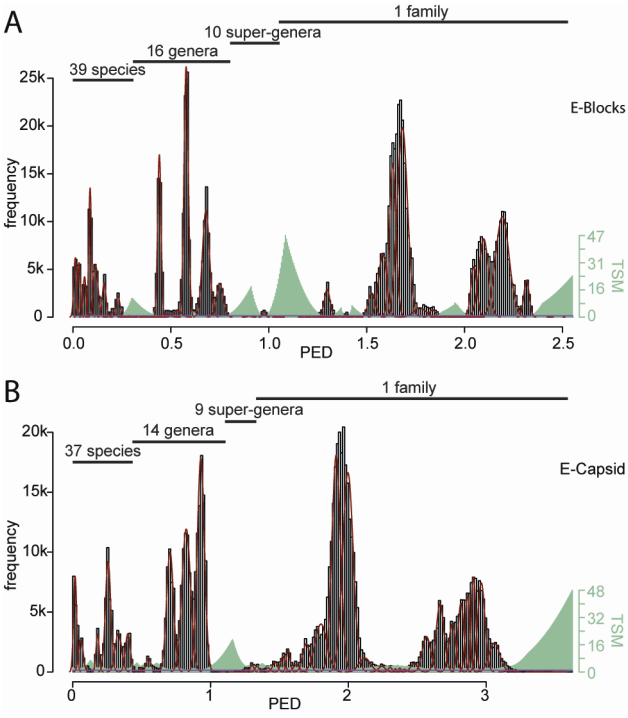


Figure 5. Impact of weakly conserved alignment regions and selection of capsid proteins on the GENETIC classification. Frequency distributions of ≈760,000 PED values formed by 1,234 picornaviruses are shown for the following evaluation data sets: a data set containing only highly conserved alignment 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.

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	BKoVα	BKoVα	-	BKoVα	BKoVα	BKoVα	-
Porcine kobuviruses	Species	BKoV	BKoVβ	BKoVβ	-	BKoVβ	BKoVβ	BKoVβ	-
HRV-C 026,NY-074,NAT001,QPM	Species	HRV-Cα	-	HRV-C	HRV-Cαβ	-	-	HRV-C	-
HRV-C 025	Species	HRV-Cβ	-	HRV-C	HRV-Cαβ	-	-	HRV-C	-
HRV-C N4,N10,NAT045	Species	HRV-Cγ	-	HRV-C	-	-	-	HRV-C	-
HRV-A	Species	HRV-A	-	HRV-A	-	-	-	HRV-A	-
HRV VR-1118,VR-1155,VR-1301	Species	HRV-Aβ	-	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	-	FMDVα	-	-	-	-	-
FMDV type SAT1,SAT2,SAT3	Species	FMDV	-	FMDVβ	-	-	-	-	-
Avian sapeloviruses	Genus	Sa	-	-	-	-	-	Saα	-
Porcine and Simian sapeloviruses	Genus	Sa	-	-	-	-	-	Saβ	-
Kobuviruses	Genus	Ko	-	KoSK	-	-	-	KoSK	-
Sali- and klasseviruses	Genus	SK	-	KoSK	-	-	-	KoSK	-
Hepatoviruses	Genus	He	-	HeTr	-	-	-	-	-
Tremoviruses	Genus	Tr	-	HeTr	-	-	-	-	-
Cardio- and senecaviruses	Supergenus	CaSe	CaSeCo	CaSeCoEr	-	-	-	-	CaSeCo
Cosaviruses	Supergenus	Co	CaSeCo	CaSeCoEr	-	-	-	-	CaSeCo
Erboviruses	Supergenus	Er	-	CaSeCoEr	-	-	-	-	-
Picornaviruses	Supergenus	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.

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-Aβ), *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): *Hepatovirus* and *Tremovirus*[314] as well as *Kobuvirus* and saliviruses[298], 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 considerably 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 HRV-Cβ 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.

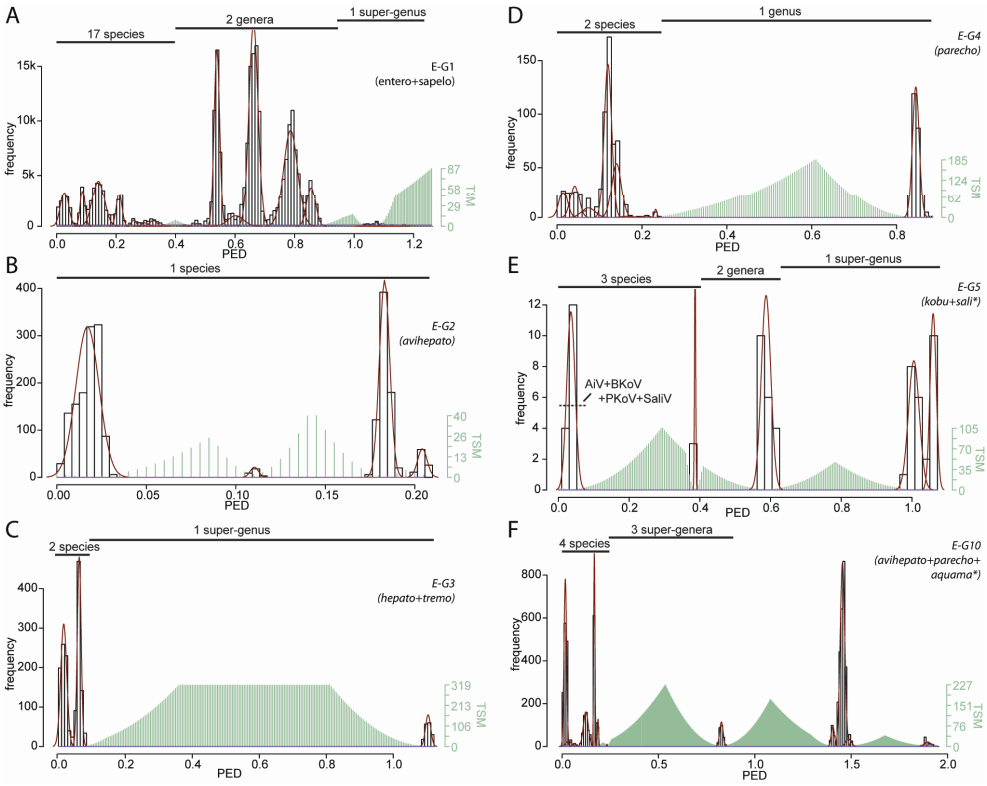


Figure **6. Reproducibility of the GENETIC classification on the species level, part one.** Frequency distributions of PED values are shown for supergenera G1 to G5 of the main data set (A to E) or a combination of three supergenera (F). PED values were compiled based on alignments covering all cluster-wide 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.

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).

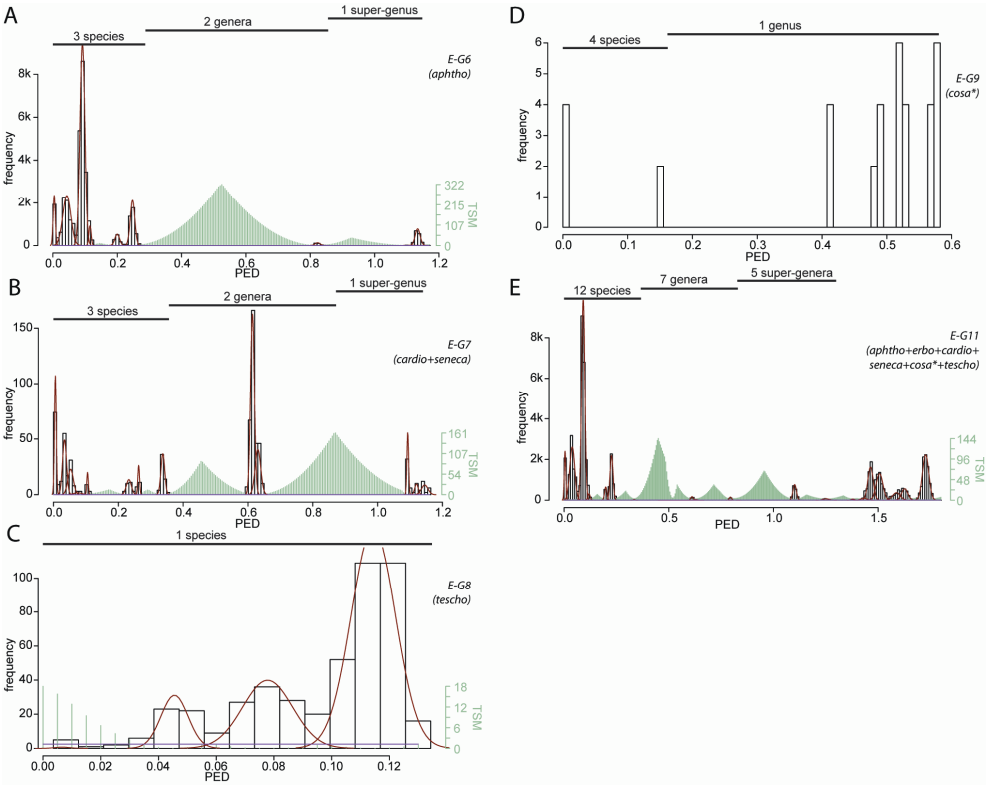


Figure 7. Reproducibility of the GENETIC classification on the species level, part two. Frequency distributions of PED values are shown for supergenera G6 to G9 of the main data set (A to D) or a combination of five supergenera (E). PED values were compiled based on alignments covering all cluster-wide 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.

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[24], 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-Aβ united

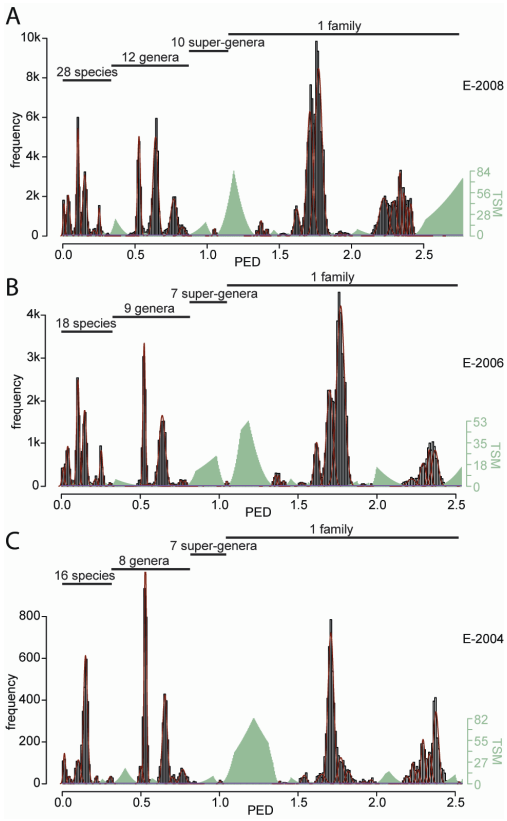


Figure 8. Impact of virus sampling on the GENETIC classification. Frequency distributions of PED values are shown for evaluation data sets formed by picornaviruses sampled until 2 years (A), 4 years (B), and 6 years (C) ago with respect to the sampling time of the main data set. The goodness-of-fit values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

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.

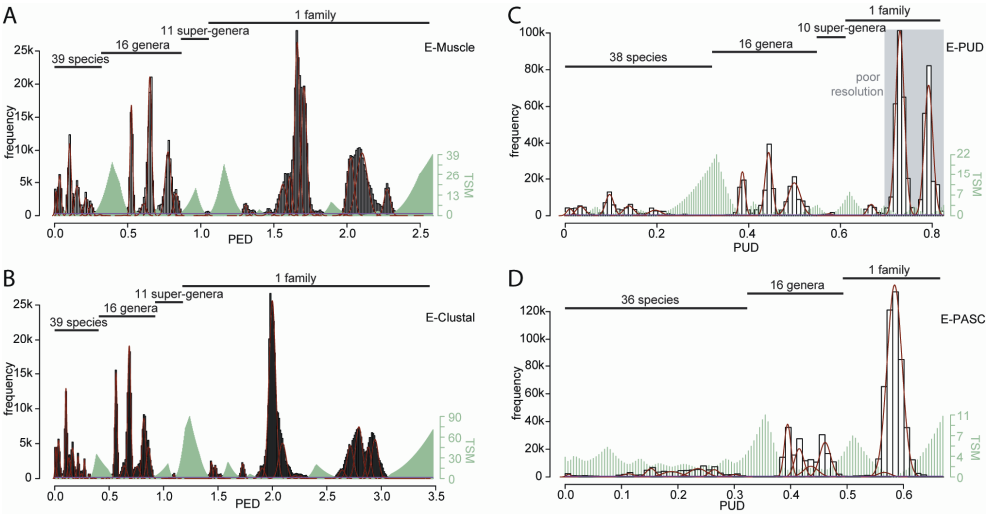


Figure 9. Impact of alignment construction and incorporation of PASC elements into the DEmARC framework on the GENETIC classification. Frequency distributions of ~760,000 PED or PUD values formed by 1,234 picornaviruses are shown for the following evaluation data sets: PEDs were calculated using the 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.

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[182,293]. 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).

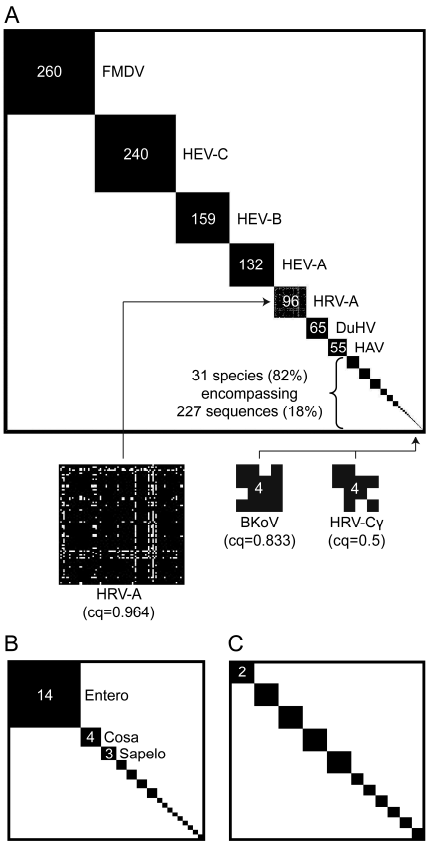


Figure 10. Sampling size of taxa and completeness of species in the GENETIC classification. (A) Shown is a binary square matrix of 1,234 viruses derived from the M-2010 PED matrix. Virus pairs whose PED does not exceed the species distance threshold are shown as black dots that form 38 species-specific 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.

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 [283]), 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 sequences, 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[117,121,211,404]. 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 descend in the entire data set. In our study, we analyzed all orthologous proteins conserved across all viruses in the data set[182]. 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 [1,38,309]; commonly a structural protein is used (see, for example, references [17,24,309,339,422,496]). 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[264].

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[139,271,311]. 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[405,425]. Outside these regions, recombination was reported exclusively for closely related viruses of the same species, although few viruses were characterized in this respect[21,39,59,238,292-294,390,405,425,428,501]. 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)[283]. 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 (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.

Table 4. Comparison of pairwise distance-based classification approaches of this study, the standard tool at NCBI, and other studies[a].

	Parameter for indicated classification approach		
Aspect	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	*a priori* defined	*a priori* 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 by other means, including rough, subjective placement and missing description (subjective).

(ii) Alignment-building procedure. We calculated pairwise distances based on a multiple alignment, which compared to widely used pairwise alignments[1,2,24,94], is expected to improve the reconstruction accuracy of orthologous relationships of sequence residues[147,191]. 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[412], 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 intragroup 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[283]. 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[443]. 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" [64,204]. 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[283]. 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[90].

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 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 program), and Leiden University Fund (Special Chair in Applied Bioinformatics in Virology).

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 paper[282], 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 importance[373], 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 reference [467]). 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 proteins[313]. 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 plan[182]. 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 host[7]. The open reading frame that encodes the polyprotein[346] 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 subfamilies[264]. 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 paper[282], 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 ICTV[264]. 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/RefSeq[36] databases were downloaded using HAYGENS[423] into the Viralis platform[183]. 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 3.52[126], 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 5.2[412]. The WAG amino acid substitution matrix[478] was applied. PEDs serve as a measure of intervirus genetic divergence.

Phylogeny reconstruction. Bayesian posterior probability trees were compiled utilizing the Beast software version 1.4.7[119]. 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 matrix[478]. 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 1.4[120]. ML trees were compiled utilizing the PhyML software version 3.0[196]. 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.

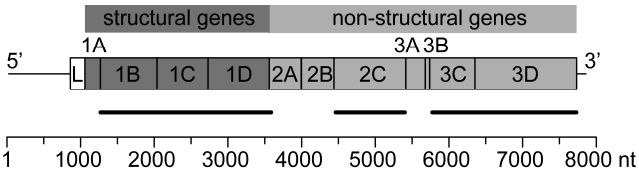


Figure 1. Picornavirus genome organization. The organization of the picornavirus genome is shown on the example of Porcine sapelovirus. Products derived after cleavage of the encoded polyprotein are indicated by rectangles and names. They include structural proteins (dark gray background) forming virus 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 divergence[282]. 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 paper[282], 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 reference [282] 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 intraspecifc 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 assignments[264]. 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 picornaviruses[282]. 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.

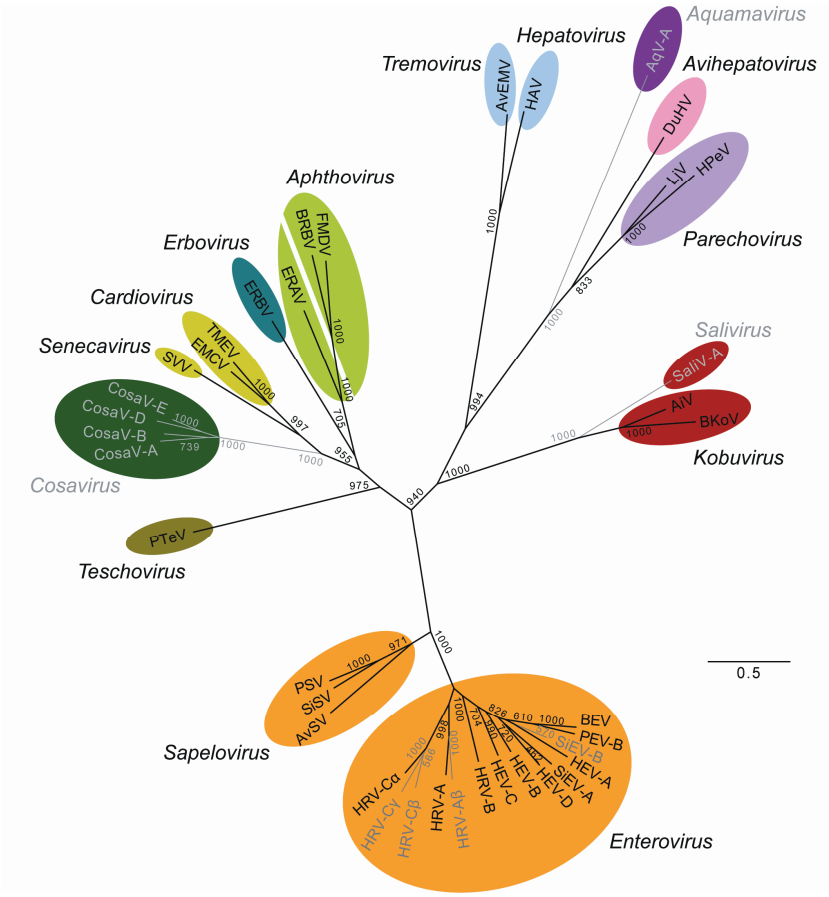


Figure 2. Phylogeny and GENETIC classification of the *Picornaviridae*. Shown is a maximum likelihood phylogeny of 38 picornaviruses representing species diversity based on the family-wide conserved proteins 1B, 1C, 1D, 2C, 3C, and 3D. A Bayesian analysis resulted in an identical tree topology (data 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 taxonomy[435], 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 Cα (HRV-Cα), Human rhinovirus Cβ (HRV-Cβ), and Human rhinovirus Cγ (HRV-Cγ) (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 thought[262].

The eight clusters encompassing exclusively novel viruses include the following: cosaviruses (4 clusters; CosaV-A, CosaV-B, CosaV-C, CosaV-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 Aβ, HRV-Aβ)[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).

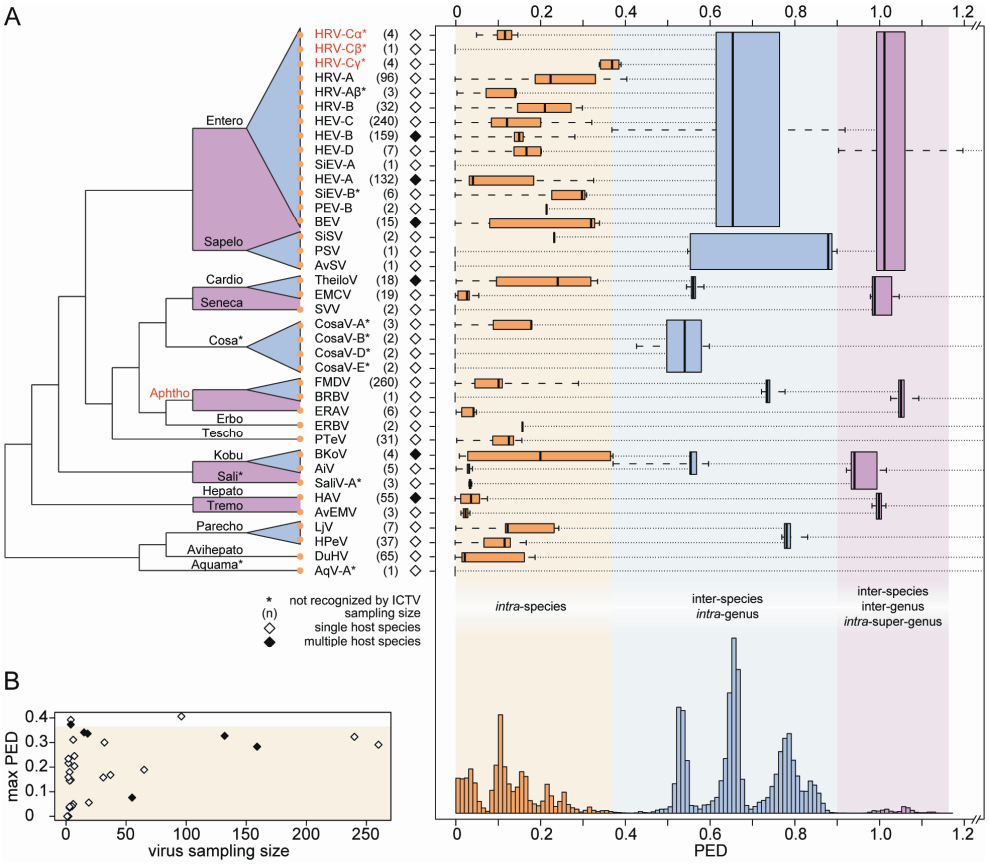


Figure 3. Intragroup genetic divergence and species sampling size. (A) Box-and-whisker graphs were used to plot distributions of distances between viruses from the same species (orange), between viruses from different species but the same genus (blue), and between viruses from different genera but the 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 reference [282]) 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.

	Difference			
Virus[a]	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-Aβ*	1
Human rhinovirus C 026, NY-074, NAT001, QPM	mm	HRV-C	HRV-Cα*	1
Human rhinovirus C 025	mm	HRV-C	HRV-Cβ*	-
Human rhinovirus C N4, N10, NAT045	mm	HRV-C	HRV-Cγ*	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-Cγ (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 sets[282]. This splitting would be in line with the original proposal by the authors who identified the porcine kobuvirus[386].

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	Supergenus	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 diver-gence 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 only[282]. 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 phylogeny[318] (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" rhinoviruses[351] (known otherwise as the cluster HRV-A2[81]) 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 reference [31]). 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 proteins[282] 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 parents[305]. 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 positions[420], 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

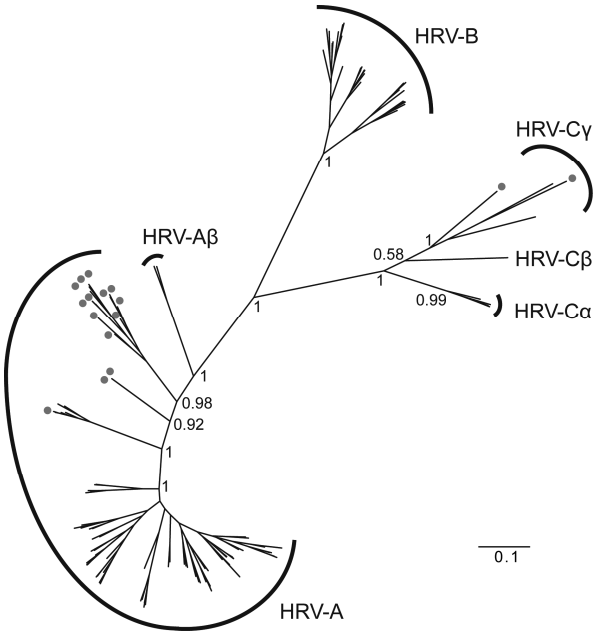


Figure 4. Phylogeny of rhinoviruses. Shown is an ML phylogeny for 140 rhinoviruses based on the family-wide conserved proteins 1B, 1C, 1D, 2C, 3C, and 3D. SH-like support values are shown for basal branching events. Species taxa recognized by the GENETIC classification are indicated (see also the Fig. 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 parts[238], 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 L[364] and 2A protein[303] or CRE[464] 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 replication[7]. 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 (supergenus) that is considered below. This result was also reproduced in classifications of two evaluation data sets[282] 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 group[264] 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 supergenus. The GENETIC classification recognizes an additional rank—provisionally called supergenus—that has no counterpart in virus taxonomy. The threshold support for

this level is the strongest overall[282], 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 intrasupergenus” 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 taxonomy[435] of the *Picornaviridae* at the species and genus levels, with notable differences concerning the recognition of only few

taxa. The observed match is nontrivial[467], 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 introduced[282].

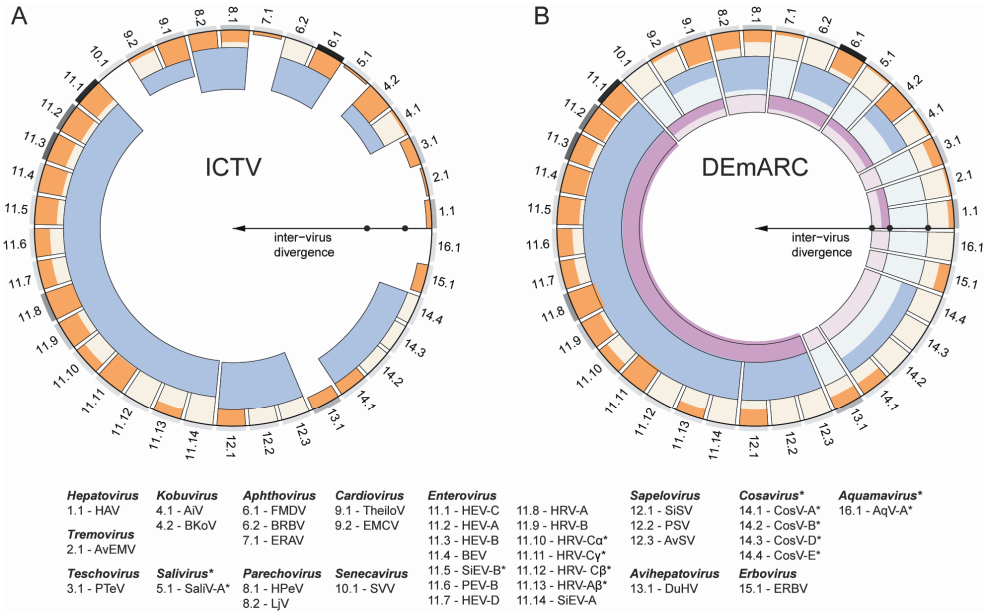


Figure 5. Taxonomy diagram and comparison of classification frameworks. Shown is a taxonomy diagram for a classification under the ICTV framework (A) and under the DEmARC framework (B). For simplicity, the GENETIC classification is visualized in both cases and supergenera are omitted for ICTV. Intervirus genetic divergence (as PED) increases linearly (arrow) from the perimeter (PED of zero) toward the center of the circle (maximum PED of 2.78). Applied distance thresholds are shown as black dots and the delimited taxa as rectangle-like shapes. Taxa are filled using the coloring scheme from Fig. 3; the three basic colors represent the species (orange), genus (blue), and supergenus (purple) levels. Each color exists in two shadings that highlight the limit on intragroup genetic divergence according to a distance threshold (soft shading) and the maximum observed intragroup genetic divergence (bright shading) of a taxon. Outside the circle, the relative density of virus sampling per species is shown as gray shadings from low (light) to high (dark) sampling, which is in the range of 1 (least sampled species) to 260 (most sampled species). For simplicity, species identities are indicated via a binary system where the first number and the second number represent the genus and the species, respectively, as defined in the common legend below the circles. (A) ICTV treats each genus independently (different heights of genus shapes) and species must conform to genus-specific distance thresholds (equal heights of species shapes only within the same genus). (B) In the DEmARC framework, taxa are treated equally at each level and they must conform to family-wide distance thresholds (equal, level-specific heights of taxon shapes). The space inside taxon shapes colored in soft shading highlights the genetic diversity that may be missed by the current picornavirus sampling, when assuming a universal, level-wide threshold that limits the actual diversity of each taxon.

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 approach[282] 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).

CHAPTER 4

Mesoniviridae: a proposed new family in the order *Nidovirales* formed by a single species of mosquito-borne viruses

Chris Lauber
John Ziebuhr
Sandra Junglen
Christian Drosten
Florian Zirkel
Phan Thi Nga
Kouichi Morita
Eric J. Snijder
Alexander E. Gorbalenya

Archives of Virology (2012) 157:1623

Abstract

Recently, two independent surveillance studies in Côte d'Ivoire and Vietnam, respectively, led to the discovery of two mosquito-borne viruses, Cavally virus and Nam Dinh virus, with genome and proteome properties typical for viruses of the order *Nidovirales*. Using a state-of- the-art approach, we show that the two insect nidoviruses are (i) sufficiently different from other nidoviruses to represent a new virus family, and (ii) related to each other closely enough to be placed in the same virus species. We propose to name this new family Mesoniviridae. *Meso* is derived from the Greek word ''mesos'' (in English ''in the middle'') and refers to the distinctive genome size of these insect nidoviruses, which is intermediate between that of the families *Arteriviridae* and *Coronaviridae*, while *ni* is an abbreviation for ''nido''. A taxonomic proposal to establish the new family Mesoniviridae, genus Alphamesonivirus, and species Alphamesonivirus 1 has been approved for consideration by the Executive Committee of the ICTV.

The order *Nidovirales*[91] includes positive-sense singlestranded RNA (ssRNA+) viruses of three families: *Arteriviridae*[140] (12.7–15.7-kb genomes; ''small-sized nidoviruses''), *Coronaviridae*[90] and *Roniviridae*[85] (26.3–31.7 kb; the last two families are jointly referred to as ''large-sized nidoviruses'')[174]. All other known ssRNA+ viruses have genome sizes below 20 kb. Recently, two closely related viruses, Cavally virus (CAVV) and Nam Dinh virus (NDiV), were discovered by two independent groups of researchers in Côte d'Ivoire in 2004 and in Vietnam in 2002, respectively[336,500]. CAVV was isolated from various mosquito species belonging to the genera *Culex*, *Aedes*, *Anopheles* and *Uranotaenia*[500]. It was most frequently found in *Culex* species, especially *Culex nebulosus*. Except for *Culex quinquefasciatus*, which circulates worldwide, the other mosquito species are endemic to Africa. NDiV was isolated from *Culex vishnui*, which is endemic to Asia, and *Culex tritaeniorhynchus*, which circulates in Asia and Africa[336], and there are indications that it may infect more mosquito species (Nga, unpublished data). Analysis of abundance patterns of 39 CAVV isolates in different habitat types along an anthropogenic disturbance gradient has indicated an increase in virus prevalence from natural to modified habitat types[243]. A significantly higher prevalence was found especially in human settlements. Analysis of habitat-specific virus diversity and ancestral state reconstruction demonstrated an origin of CAVV in a pristine rainforest with subsequent spread into agriculture and human settlements[500]. Notably, it was shown for the first time that virus diversity decreased and prevalence increased during the process of emergence from a pristine rainforest habitat into surrounding areas of less host biodiversity due to anthropogenic modification[500]. Both viruses were propagated in *Aedes albopictus* cells and characterized using different techniques. A number of common properties place CAVV and NDiV in the order *Nidovirales*. These properties include (i) the genome organization with multiple open reading frames (ORFs), (ii) the predicted proteomes (Fig. 1), (iii) the production of enveloped, spherical virions, and (iv) the synthesis of genome-length and subgenome-length viral RNAs in infected cells[336,500]. Particularly, the two viruses were found to encode key molecular markers characteristic of all nidoviruses: a 3C-like main protease (3CLpro, also known as Mpro) flanked by two transmembrane (tM) domains encoded in replicase ORF1a, as well as an RNA-dependent RNA polymerase (RdRp) and a combination of a Zn-binding module (Zm) fused with a superfamily 1 helicase (HEL1) encoded in ORF1b. As in other nidovirus genomes, ORFs 1a and 1b were found to overlap by a few nucleotides in both CAVV and NDiV. The ORF1a/1b overlap region includes a putative -1 ribosomal frameshift site (RFS) that is expected to direct the translation of ORF1b by a fraction of the ribosomes that start translation at the ORF1a initiation codon. Thus, a frameshift just upstream of the ORF1a termination codon mediates the production of a C-terminally extended polyprotein jointly encoded by ORF1a and ORF1b. Combined, these markers form the characteristic nidovirus constellation: tM-3CLpro-tM_RFS_RdRp_Zm-HEL1 (Fig. 1)[91,174]. Likewise, virion proteins are encoded in ORFs that are located downstream of ORF1b and expressed from a set of subgenomic mRNAs.

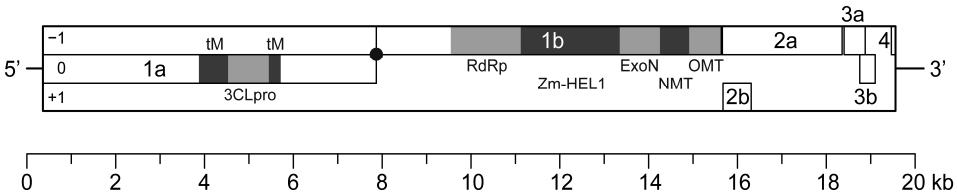


Figure 1. Genome organization of mesoniviruses. The coding and 5'- and 3'-untranslated regions of the genome are represented, respectively, by the outer rectangle and horizontal lines. ORFs are shown as open rectangles and are arranged in three reading frames (-1, 0, +1) relative to that of ORF1a. ORF1a- and ORF1b-encoded protein domains identified by bioinformatics analysis (see ref. [336]) are highlighted in grey. The predicted location of -1 ribosomal frameshift signals are indicated by a black dot. The genome organization is shown for NDiV but is virtually identical to that of CAVV except for the reading frame of some ORFs (see Table 1).

No similarities were found between the (putative) structural proteins of CAVV and NDiV and those of other nidoviruses[336,500]. The most distinctive molecular characteristic of CAVV and NDiV, however, is the ~20-kb genome size, that is intermediate between the size ranges of small-sized and large-sized nidovirus genomes. Consequently, each of the two viruses has been proposed to prototype a new nidovirus family[336,500].

In this study, we compared the genomes of CAVV (GenBank accession number HM746600) and NDiV (GenBank accession number DQ458789) to assess their relationship and use this insight for taxonomic classification of these viruses. To date, only very limited biological information is available for CAVV and NDiV (see above), and in general, biological properties may be affected profoundly by a few changes in the genome. In view of these considerations and in line with the accepted taxonomic approach to viruses of the family *Coronaviridae*[90], comparative sequence analysis was considered the most reliable basis for classification. The overall similarity between the CAVV and NDiV genomes was found to be strikingly high: nearly identical sizes (20,187 and 20,192 nt, respectively), conservation of ORFs with sequence identities ranging from 87.8 to 96.1% at the amino acid level and from 88.3 to 93.7% at the nucleotide level (Table 1). Given this high similarity, prior assignments of domains and genetic signals were cross-checked to produce a unified description.

There was complete agreement between the two studies[336,500] on the mapping of all nidovirus-wide conserved domains in CAVV and NDiV, as well as on the identification of GGAUUUU as a plausible slippery sequence in RFS (see above). Additionally, our analysis showed that the NDiV-based assignment[336] of 3'-to-5' exoribonuclease (ExoN) and 2'-O-methyltransferase (OMT), two replicative domains characteristic for large-sized nidoviruses[174], and N7-methyltransferase (NMT)[73] in ORF1b extends to CAVV. Likewise, CAVV may lack a uridylate-specific endonuclease (NendoU), as has previously been observed for NDiV[336]. The synthesis of subgenomic RNAs from which ORFs 2a to 4 are predicted to be expressed appears to be controlled by transcription-regulating sequences (TRSs)[136,353,408] identified upstream of ORF2a/2b, ORF3a and ORF4 (collectively designated

as body TRSs). Other putative TRSs were identified downstream of the leader region located at the 5'-end of the viral genome[336,500]. Unique among nidoviruses, NDiV and CAVV may use different leader TRSs during the synthesis of different subgenomic RNAs, although further analysis is required to clarify the basis for some discrepancies between the TRS assignment in NDiV and CAVV. Also, it remains to be shown why the high sequence conservation of virion proteins of the two viruses (Table 1) was not manifested in the morphology observed upon EM analysis of virus particles[336,500]. In this respect, it may be relevant that Zirkel et al.[500] noticed two types of particles in CAVV-infected cells, one of which carried club-shaped surface projections compatible with viral glycoproteins. This latter type of particles was also observed in infected cell culture supernatant. Ultimately, the origin of the particles of both types, and their relationship to the particles isolated from the medium of NDiV-infected C6/36 cells by Nga et al.[336] should be revealed by future research efforts.

Furthermore, we evaluated the phylogenetic position of CAVV and NDiV in relation to other nidoviruses. We conducted a phylogenetic analysis as described in ref. [336]. The study indicates that CAVV and NDiV consistently, albeit very distantly, cluster with viruses of the family *Roniviridae*, the only other known nidoviruses infecting invertebrates (Fig. 2). Quantitatively, this Bayesian posterior probability phylogeny illustrates that CAVV and NDiV form a deeply rooted lineage in the nidovirus tree with an evolutionary divergence from other nidoviruses comparable to that separating viruses of the families *Coronaviridae* and *Roniviridae* (Fig. 2). Together, these characteristics of CAVV and NDiV (insect host, intermediate genome size, deeply rooted phylogenetic lineage) provide a compelling basis for the creation of a new nidovirus family. We propose to name this new family Mesoniviridae, where *meso* is derived from the Greek word ''mesos'' (in English ''middle'' or ''in the middle'') and refers to a key distinctive characteristic of these viruses, namely their intermediate-sized genomes. The second component of the acronym, *ni*, refers to nidoviruses, as has been done previously for ro*ni*viruses[84] and bafi*ni*viruses[414].

Table 1. Comparison of ORFs in the genome of NDiV and CAVV.

	Length [nt]		Frame[a]		Identity [%][b]		
	NDiV	CAVV	NDiV	CAVV	nt	aa	Predicted protein
ORF1a	7509	7497	0	0	88.3	90.0	Polyprotein 1a
ORF1b	7587	7587	-1	-1	92.6	96.1	1b part of polyprotein 1ab
ORF2a	2697	2700	-1	-1	90.7	87.5	Spike
ORF2b	636	642	+1	+1	88.8	90.2	Nucleocapsid
ORF3a	474	474	-1	+1	91.1	93.0	Membrane
ORF3b	348	348	0	-1	93.7	90.5	Membrane
ORF4	135	147	+1	-1	89.9	87.8	Unknown

ORF designations according to Table 2 in ref. [500]

[a] Reading frame relative to that of ORF1a

[b] Pairwise nucleotide (nt) and amino acid (aa) sequence identity between NDiV and CAVV

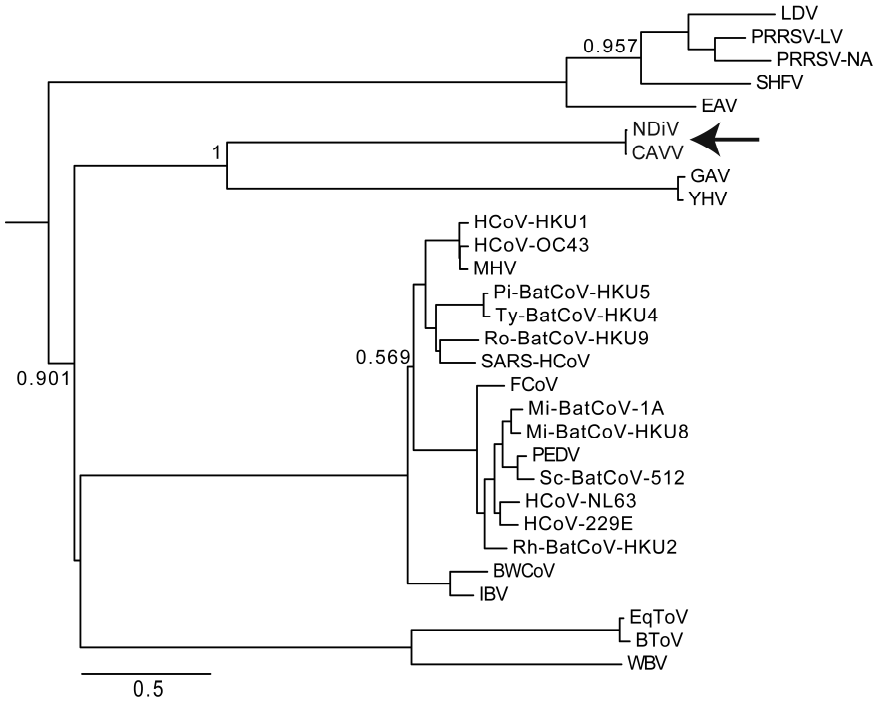


Figure 2. Phylogenetic position of CAVV and NDiV. To infer phylogenetic relationships of Nam Dinh virus isolate 02VN178 (NDiV), Cavally virus isolate C79 (CAVV) (arrow) and other nidoviruses, a partially constrained tree was calculated using a concatenated alignment of the three nidovirus-wide conserved domains and a set of viruses representing currently recognized species. The alignment was produced with Muscle version 3.52[126] in the Viralis platform[183], and the phylogenetic analysis was performed using BEAST version 1.4.7[119]. For further details, see ref. [336]. Numbers indicate posterior probability support values (on a scale from 0 to 1); all internal nodes for which no support value is provided have been fixed in the analysis based on prior analyses of nidovirus subsets (data not shown). The scale bars represent the average number of substitutions per amino acid position. The tree was rooted on the arterivirus branch. Virus names and GenBank/Refseq accession numbers: lactate dehydrogenase- elevating virus (LDV; U15146), porcine respiratory and reproductive syndrome virus European type (PRRSV-LV; M96262), porcine respiratory and reproductive syndrome virus North American type (PRRSV-NA; AF176348), simian hemorrhagic fever virus (SHFV; NC_003092), equine arteritis virus (EAV; AY349167), Nam Dinh virus (NDiV; DQ458789), Cavally virus (CAVV; HM746600), gill-associated virus (GAV; AF227196), yellow head virus (YHV; EU487200), human coronavirus HKU1 (HCoV-HKU1; AY884001), human coronavirus OC43 (HCoV-OC43; AY585228), mouse hepatitis virus (MHV; AY700211), Pipistrellus bat coronavirus HKU5 (Pi-BatCoV-HKU5; EF065509), Tylonycteris bat coronavirus HKU4 (Ty-BatCoV-HKU4; EF065505), Rousettus bat coronavirus HKU9 (Ro-BatCoV-HKU9; EF065513), SARS coronavirus (SARS-HCoV; AY345988), feline coronavirus (FCoV; NC_007025), Miniopterus bat coronavirus 1A (Mi-BatCoV-1A; NC_010437), Miniopterus bat coronavirus HKU8 (Mi-BatCoV-HKU8; NC_010438), porcine epidemic diarrhoea virus (PEDV; NC_003436), Scotophilus bat coronavirus 512 (Sc-BatCoV-512; DQ648858), human coronavirus NL63 (HCoV-NL63; DQ445911), human coronavirus 229E (HCoV-229E; NC_002645), Rhinolophus bat coronavirus HKU2 (Rh-BatCoV-HKU2; NC_009988), beluga whale coronavirus SW1 (BWCoV; EU111742), avian infectious bronchitis virus (IBV; NC_001451), equine torovirus (EToV; X52374), bovine torovirus (BToV; NC_007447), white bream virus (WBV; NC_008516).

Next, we sought to establish species demarcation criteria to decide whether CAVV and NDiV prototype separate species or belong to a single species. Commonly, this question cannot be answered (reliably) on the basis of only two full genome sequences and otherwise very limited biological data. To solve this dilemma, we exploited information available for other nidoviruses in our analysis. In order to evaluate the genetic similarity between CAVV and NDiV in the context of sequence divergence of lineages representing previously established nidovirus species, we applied a state-of-the-art framework for a genetics-based classification[282]. This recently introduced classification approach has been shown to recover and refine the taxonomy of picornaviruses[283], and it was also used to revise the taxonomy of coronaviruses extensively (Lauber & Gorbalenya, in preparation)[90]. In addition to CAVV and NDiV, a representative set of 152 large-sized nidoviruses was included in the analysis. Two sets of proteins were used: the first included proteins conserved in all nidoviruses (3CLpro, RdRp, HEL1) (dataset D1), while the second set additionally included ExoN and OMT, which are conserved in large-sized nidoviruses and CAVV/NDiV (dataset D2). For both datasets a concatenated, multiple amino acid alignment was produced, which formed the basis for compiling pairwise evolutionary distances (PEDs) between all pairs of viruses (Fig. 3ab; for details see ref. [282]). It was found that the PED separating CAVV and NDiV is within the range of intra-species virus divergence in the families *Coronaviridae* and *Roniviridae* for both datasets (Fig. 3cd). Specifically, CAVV and NDiV show a distance (0.016 and 0.029 for D1 and D2, respectively) that is below the genetic divergence of members of several established nidovirus species (maximum of 0.032 and 0.037 for D1 and D2, respectively). For both datasets, these viruses include gill-associated virus and yellow head virus (species *Gill-associated virus*, family *Roniviridae*)[85] and the coronaviruses feline coronavirus, transmissible gastroenteritis virus, and porcine respiratory coronavirus (*species Alphacoronavirus 1*), IBV (species *Avian coronavirus*), murine hepatitis virus (species *Murine coronavirus*), and Rousettus bat coronavirus HKU9 (species *Rousettus bat coronavirus HKU9*)[90]. For the dataset comprising the three nidovirus-wide conserved proteins (Fig. 3ac), Miniopterus bat coronavirus 1 also showed a maximum genetic divergence exceeding that of the CAVV-NDiV pair. Together, these observations show that CAVV and NDiV belong to the same species, representing a single genus in the family. We propose to name this genus Alphamesonivirus and the species Alphamesonivirus 1, thereby following a naming convention recently applied to the subfamily *Coronavirinae*[90], which is expected to facilitate the accommodation of future expansions of the family. A taxonomic proposal for family, genus, and species recognition has been available on-line at the ICTV website (http://talk.ictvonline.org/files/proposals/taxonomy_proposals_ invertebrate1/ m/default.aspx) since August 2011. It has been approved by the chairs of the ICTV *Arteriviridae*, *Coronaviridae*, and *Roniviridae* Study Groups and the Executive Committee of the ICTV, and will be considered again at the next EC-ICTV meeting, to be held in Leuven, Belgium, in July 2012.

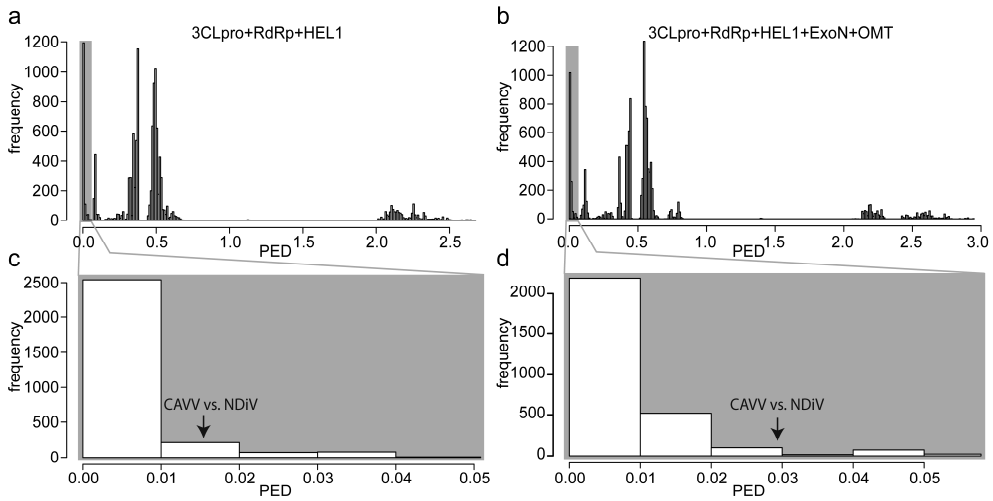


Figure 3. Evolutionary distance between CAVV and NDiV in relation to intra-species genetic divergence in large-sized nidoviruses. Multiple amino acid alignments for 154 nidoviruses with large genomes (all major nidovirus lineages except arteriviruses) comprising three nidovirus-wide conserved protein domains (**a**, **c**) or five domains conserved in all large-sized nidoviruses (**b**, **d**) were used to compile pairwise evolutionary distances (PEDs) between all virus pairs. These distances are shown as frequency distributions (**a**, **b**), and zoom-ins on small distances are provided (**c**, **d**). The PED between CAVV and NDiV (indicated by the arrow) is well within the intra-species distance range of other nidoviruses. Several currently recognized nidovirus species show a maximum genetic divergence larger than that of the CAVV-NDiV pair (see text).

The recognition of CAVV and NDiV as a single virus species can be contrasted with the detection of these viruses in many mosquito host species and their spread to different continents (Africa and Asia, respectively)[336,500]. The underlying mechanisms of this broad dispersal are unknown but might include the crossing of the host species barrier rather than virus-host cospeciation. Further research, including the characterization of biological properties of CAVV and NDiV and the extension of surveillance studies to other regions of the world, is needed to understand the ecology, host tropism and medical and/or economic relevance of mesoniviruses.

Acknowledgments

We thank Kay Faaberg, Jeff Cowley, and Raoul de Groot for reviewing the taxonomic proposal to establish the family Mesoniviridae that was used to draft this publication. Expert administration of the Viralis platform by Igor Sidorov, Alexander Kravchenko and Dmitry Samborskiy is acknowledged. This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under the programs SILVER (grant

agreement no. 260644), EMPERIE (grant agreement number 223498), and EVA (grant agreement number 228292), and the Deutsche Forschungsgemeinschaft (grant agreement number DR772/3-1), LUMC MoBiLe program, Leiden University Fund, the BonFor programme of the University of Bonn (Grant agreement number O-156-0006), and the Program of Japan Initiative for Global Research Network on Infectious Diseases (J-GRID), MEXT, Japan.

Open Access

CHAPTER 5

Discovery of the First Insect Nidovirus, a Missing Evolutionary Link in the Emergence of the Largest RNA Virus Genomes

Phan Thi Nga[#]
Maria del Carmen Parquet[#]
Chris Lauber[#]
Manmohan Parida
Takeshi Nabeshima
Fuxun Yu
Nguyen Thanh Thuy
Shingo Inoue
Takashi Ito
Kenta Okamoto
Akitoyo Ichinose
Eric J. Snijder
Kouichi Morita
Alexander E. Gorbalenya

[#] joint first authors

PLoS Pathogens (2011) 7:e1002215
DOI:10.1371/journal.ppat.1002215

Abstract

Nidoviruses with large genomes (26.3–31.7 kb; 'large nidoviruses'), including *Coronaviridae* and *Roniviridae*, are the most complex positive-sense single-stranded RNA (ssRNA+) viruses. Based on genome size, they are far separated from all other ssRNA+ viruses (below 19.6 kb), including the distantly related *Arteriviridae* (12.7–15.7 kb; 'small nidoviruses'). Exceptionally for ssRNA+ viruses, large nidoviruses encode a 3′-5′exoribonuclease (ExoN) that was implicated in controlling RNA replication fidelity. Its acquisition may have given rise to the ancestor of large nidoviruses, a hypothesis for which we here provide evolutionary support using comparative genomics involving the newly discovered first insect-borne nidovirus. This Nam Dinh virus (NDiV), named after a Vietnamese province, was isolated from mosquitoes and is yet to be linked to any pathology. The genome of this enveloped 60–80 nm virus is 20,192 nt and has a nidovirus-like polycistronic organization including two large, partially overlapping open reading frames (ORF) 1a and 1b followed by several smaller 3′-proximal ORFs. Peptide sequencing assigned three virion proteins to ORFs 2a, 2b, and 3, which are expressed from two 3′-coterminal subgenomic RNAs. The NDiV ORF1a/ORF1b frameshifting signal and various replicative proteins were tentatively mapped to canonical positions in the nidovirus genome. They include six nidovirus-wide conserved replicase domains, as well as the ExoN and 2′-O-methyltransferase that are specific to large nidoviruses. NDiV ORF1b also encodes a putative N7-methyltransferase, identified in a subset of large nidoviruses, but not the uridylate-specific endonuclease that – in deviation from the current paradigm - is present exclusively in the currently known vertebrate nidoviruses. Rooted phylogenetic inference by Bayesian and Maximum Likelihood methods indicates that NDiV clusters with roniviruses and that its branch diverged from large nidoviruses early after they split from small nidoviruses. Together these characteristics identify NDiV as the prototype of a new nidovirus family and a missing link in the transition from small to large nidoviruses.

Author Summary

Research in virology is driven towards the characterization of a limited number of socioeconomically important pathogens, mostly those infecting humans. Yet, characterization of other viruses may advance our understanding of these topical pathogens and the fundamentals of virology. Here we describe the discovery of a virus of unknown clinical relevance that has many remarkable features. The virus was coined Nam Dinh virus (NDiV) after a Vietnamese province. It is a mosquito-borne virus with a 20.2 kilobase genome, the largest among non-segmented single-stranded RNA viruses of insects. Employing bioinformatics tools, we show that NDiV prototypes a new family and is a missing evolutionary link connecting the distantly related nidoviruses with small and large genomes, including important and diverse pathogens such as porcine respiratory and reproductive syndrome virus (~15-kilobase genome) and SARS coronavirus (~30 kilobases), respectively. NDiV and large nidoviruses form a phylogenetic cluster and share a set of core replicative enzymes. They exclusively encode an exoribonuclease that presumably controls replication fidelity. Its acquisition may have promoted the emergence of viruses with single-stranded RNA genomes larger than ~20 kilobases. This study highlights the benefits of broad virus discovery efforts for fundamental and applied research.

Introduction

Viruses employing positive-sense, single-stranded RNA genomes (ssRNA+) form the most abundant class and its members are known to infect all types of hosts except *Archaea*. They have evolved genome sizes in the range of ~3.0 to 31.6 kb (Fig. 1). This size range is the largest among those of the different classes of RNA viruses, although it is small compared to those of DNA viruses and cellular organisms. These profound genome size differences between RNA and DNA life forms are inversely correlated with mutation rates, which are highest in RNA viruses, thought due to the lack of proofreading during replication[117,236,404].

Recently, the molecular basis of the relation between RNA virus genome sizes and mutation rates has been revisited in studies of nidoviruses with large genomes ("large nidoviruses"). These viruses, with genomes of 26.3 to 31.6 kb, include the *Coronaviridae* and *Roniviridae* families and are at the upper end of the RNA virus genome size range[174]. They are uniquely separated from other ssRNA+ viruses (3.0–19.6 kb genomes), including the distantly related *Arteriviridae* family (12.7–15.7 kb genomes; "small nidoviruses") with which they form the order *Nidovirales*[174,316,360]. The order includes five major lineages of viruses that infect vertebrate and invertebrate hosts. Their complex genetic architecture includes multiple open reading frames (ORFs) that are expressed by region-specific mechanisms. The first two regions are formed by the two 5′-most and partially overlapping ORFs, ORF1a and ORF1b, which are translated from the genomic RNA to produce

polyproteins 1a (pp1a) and pp1ab. The expression level of ORF1b is downregulated relative to that of ORF1a by the use of the ORF1a/1b ribosomal frameshifting signal[54,366]. Both pp1a and pp1ab are autoproteolytically processed by ORF1a-encoded proteases to yield numerous products that control genome expression and replication[498]. The third, 3′-located region of the nidovirus genome includes multiple smaller ORFs (3′ORFs), although the number of these ORFs varies considerably among nidoviruses. These genes are expressed from 3′-coterminal subgenomic mRNAs to produce the structural proteins incorporated into the enveloped nidovirus particles and, optionally, other proteins modulating virus-host interactions[53,136,353,408]. With the exception of a few nidoviruses, the subgenomic and genomic mRNAs are also 5′-coterminal. A mechanism of discontinuous negative-stranded RNA synthesis, yielding the templates for subgenomic mRNA production, is thought to control this mosaic structure of nidovirus mRNAs. The synthesis of subgenome-length negative stranded RNAs is guided by short transcription-regulating sequences (TRSs) – located in the common “leader sequence” (near the genomic 5′ end) and in each “mRNA body” (upstream of the expressed ORFs) - that share a conserved core sequence and flank the genome region that is not present in the respective subgenomic mRNAs.

The nidovirus ORF1b encodes key replicative enzymes whose number and type vary between the major nidovirus lineages. They invariably include an RNA-dependent RNA polymerase (RdRp) and a superfamily 1 helicase (HEL1)[169], which are most common in other RNA viruses, and several other RNA-processing enzymes that are either unique to nidoviruses (uridylate-specific endonuclease (NendoU) and 3′-to-5′exoribonuclease (ExoN)) or rarely found outside nidoviruses (2′-O-methyltransferase (OMT);[174]). Among these enzymes, the ExoN domain has properties that are most relevant for understanding the relation between genome size and mutation rate in RNA viruses.

Bioinformatics-based analysis originally identified the ExoN domain only in the genomes of large nidoviruses and mapped it in the vicinity of HEL1, a key replicative enzyme[432]. It also revealed a distant relationship between ExoN and a cellular DNA-proofreading enzyme. Based on these observations, nidoviruses were proposed to have acquired ExoN to control the replication fidelity of their expanding genome[432]. The enzymatic activities of ExoN were subsequently verified and detailed in biochemical studies[72,323]. Likewise, and in line with the expectations, ExoN-inactivating mutations were shown to decrease RNA replication fidelity by ~15–20 fold in two coronaviruses, mouse hepatitis virus (MHV) and SARS coronavirus (SARS-CoV), while only modestly affecting virus viability[123,124]. These results strongly support a critical role of ExoN in the control of replication fidelity of large nidoviruses, although more mechanistic insight is clearly required before the current paradigm connecting RNA virus mutation rates and genome size control could be definitively revised to include proof-reading during the replication of large RNA genomes[99].

Major advancements toward this goal are expected to come from studies of the structure and function of ExoN, which aim to elucidate the molecular mechanism of its

action. In addition, genomics studies could contribute to this quest by providing insights into the role of ExoN in RNA virus evolution. Accordingly, if ExoN was acquired to ensure the expansion of RNA genomes beyond a certain size, we may expect (i) a genome size threshold that separates RNA viruses with and without ExoN; (ii) all nidoviruses with genome sizes above this threshold to encode ExoN; and (iii) no other domain than ExoN to correlate, functionally and phyletically, with genome size control in large nidoviruses.

In this respect, the characterization of nidoviruses with a genome size in the gap that currently separates small and large nidoviruses should, in theory, be particularly insightful. However, whether these viruses actually exist has thus far remained an open question. Three considerations suggest that if nidoviruses with intermediate-sized genomes ever evolved they may already have gone extinct. First, it is recognized that the evolution of RNA viruses is characterized by a high birth-death rate and the extinction of numerous virus lineages, resulting in the fast turnover of species[217]. Secondly, the genome size gap between large nidoviruses and all other known ssRNA+ viruses has existed without exception since genome sequencing began in the 1980s. As of the late 1980s, this gap has been bordered by closteroviruses (from the bottom) and nidoviruses (from the top) (Fig. 1). Likewise and thirdly, all nidovirus genomes sequenced to date have sizes that are similar to either IBV (27,600 nt)[49] or EAV (12,700 nt)[98], which were the first fully sequenced coronavirus and arterivirus genomes, respectively. The evident under-representation of RNA viruses with relatively large genomes is even more striking in the light of the continuous flow of newly identified ssRNA+ viruses with smaller genome sizes[32] (Fig. 1).

In sharp contrast to these considerations and prior observations, we here report the discovery of a nidovirus with a genome size that is intermediate between those of small and large nidoviruses. This elusive and precious evolutionary link is an insect-borne virus with the largest ssRNA+ genome for any insect virus known to date. Comparative genome analyses involving this newly identified virus provide evolutionary evidence for the acquisition of the ExoN domain by a nidovirus (ancestor) with a genome size in the range of ~16–20 kb. This range appears to define the size limit for the expansion of ssRNA+ virus genomes, which may be achieved in evolution without the recruitment of a specialized enzyme that controls replication fidelity. Furthermore, we found that two other replicative enzymes, N7-methyltransferase (NMT) and NendoU, are not encoded by toroviruses and invertebrate nidoviruses, respectively, indicating that they may contribute “optional activities” for the nidovirus replication machinery. Together our results highlight the broad benefits of virus discovery efforts applied to mosquitoes.

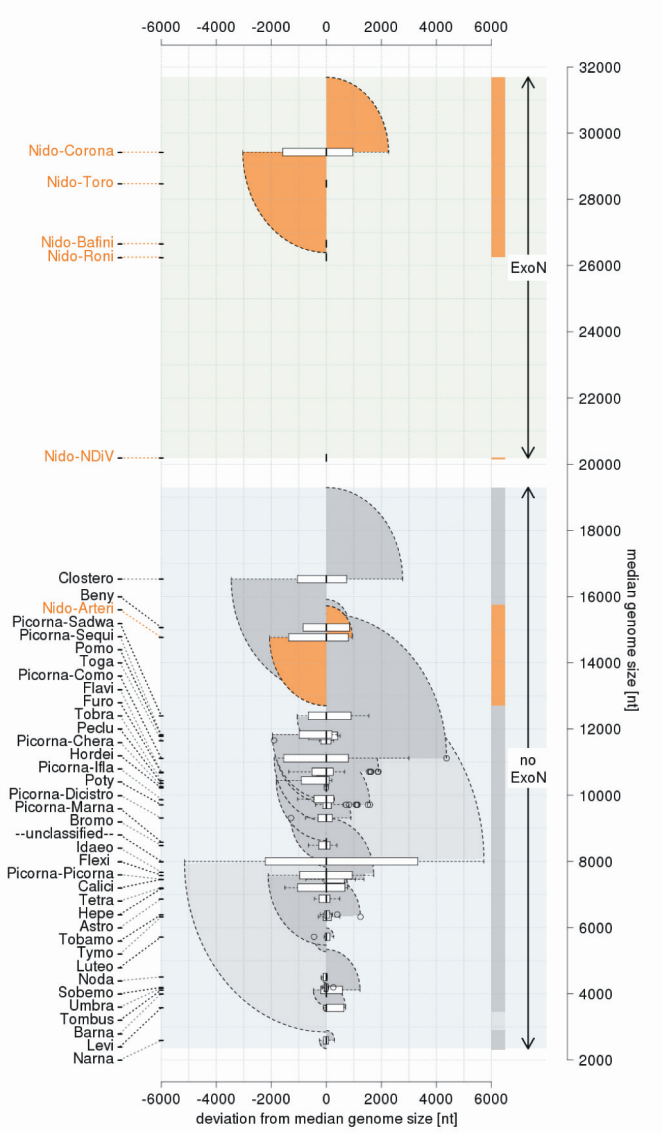


Figure 1. Distribution of positive-sense single-stranded RNA virus genome sizes. The *Coronaviridae* is split into the corona- and toro-/bafinivirus groups. Prefix Nido- and Picorna- are for *Nidovirales* and *Picornavirales*, respectively. Group specific box-whisker plots are aligned along the X-axis by their medians (bold line) normalized to zero. The box spans from the first to third quartile; the whiskers (dashed lines) are <1.5 times the inter-quartile range; outliers are circles. Families/groups are ranked by median genome size along Y-axis. Genome size ranges are colored by semi-circles: nidoviruses (dark orange), other classified (dark grey) and unclassified viruses (light grey). Three non-overlapping zones regarding the presence of the exoribonuclease (ExoN) are highlighted in the genome size distribution (from top to bottom): ExoN-encoding large nidoviruses and NDiV (light green); in-between not-sampled size zone (white); ExoN-lacking ssRNA+ viruses (light blue).

Results

Virus field study. In Vietnam, between 2,000 and 3,000 cases of acute encephalitis syndrome (AES) are reported annually, of which about 40% are confirmed to be associated with Japanese encephalitis virus (JEV). The etiological agent(s) in the other 60% of cases remains unknown[335], but they share demographic characteristics and seasonality with the JEV cases. Hence, the involvement of other arboviruses in non-JE AES was postulated and the virus described in this paper was identified in search of such pathogens, which may infect both humans and mosquitoes.

During continued JEV surveillance between September 2001 and December 2003, 359 pools containing one of six mosquito species (see Materials and Methods) were collected indoors in Northern and Central Vietnam at one- to three-month intervals. The study areas included Hanoi and other cities located in the provinces of Ha Nam (Chuyenngoan, Mocbac), Ha Tay (Catque, Phuman and Chuongmy), Nam Dinh, and Quang Binh (Fig. 2). The majority of Catque inhabitants are farmers who cultivate rice in watered paddy fields and raise pigs. Phuman and Quangbinh, however, are highlands.

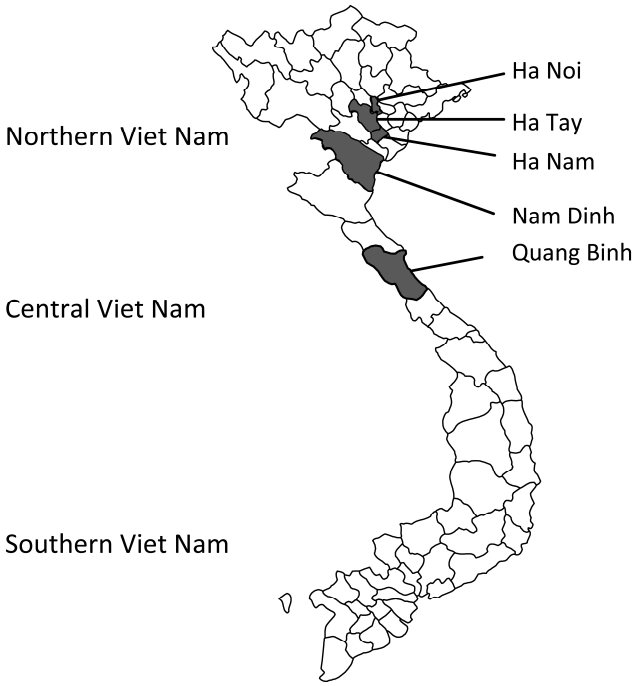


Figure 2. Map of the Vietnam provinces, where mosquito surveillance was conducted between 2001 and 2003.

Discovery of the first mosquito-borne nidovirus having an intermediate genome size. Mosquito pools were tested for the presence of viruses using infection of different cell lines as a read-out assay. Homogenates that were prepared from some pools containing *Culex tritaeniorhynchus* and *Culex gelidus* induced cytopathic effects in the C6/36 mosquito cell line. Most of these were attributed to JEV (24 different strains; data not shown), but for 10 specimens a routine laboratory screening for JEV and other circulating flaviviruses (such as Dengue and West Nile viruses) by RT-PCR and/or serology yielded negative results.

Subsequently, infected culture fluid (ICF) from cells infected with unknown agents were analyzed by electron microscopy, which revealed an enveloped virus with a diameter of 60–80 nm (Fig. 3A). This virus was named Nam Dinh virus (NDiV), after the geographic locality of its first apparent isolation, although this origin could not be confirmed later on. However, for historical reasons, this name was retained for all subsequent isolates, and the analysis of one of those (02VN178) is described here. NDiV was identified in four mosquito pools, two from *Culex vishnui* and two from *Culex tritaeniorhynchus*, collected in two other provinces of Vietnam (Table 1). PCR amplification using virus-specific primers to an ORF1b region (see below) was employed to verify the presence of NDiV in the mosquito samples, but to date no other insects have been probed for the presence of the virus. It also remains to be investigated whether NDiV causes disease in susceptible hosts and whether it may infect humans.

Purified NDiV was used for virion protein analysis (Fig. 3B) and genome sequencing (Fig. S1; Materials and Methods). *In silico* translation of the unsegmented, 20,192 nt-long NDiV genome (GenBank accession number DQ458789) indicated that it contained at least six ORFs: ORF1a (nt 361–7869), ORF1b (7830–15635), ORF2a (15660–18356), ORF2b (15674–16309), ORF3 (18402–18875) and ORF4 (18754–19101) (Fig. 3D). The region encompassing ORFs 3 and 4 also contains a few smaller potential ORFs. The coding region of the genome is flanked by a 5′-untranslated region (UTR) (1–360) and a 3′-UTR (19102–20192), with the latter being followed by a poly(A) tail. The 5′-UTR includes two AUG codons indicating that translation initiation for ORF1a/ORF1b is likely mediated by another mechanism than ribosomal scanning. Three pairs of ORFs (1a–1b, 2a–2b, and 3–4) overlap to variable degrees; particularly, ORF1b overlaps ORF1a in the −1 frame (Fig. 3D; see also below). Overall, these results showed that NDiV is an insect-borne ssRNA+ virus with the largest genome known so far - twice the size of the next largest one, which is the genome of the *Iflavirus* Brevicoryne brassicae picorna-like virus[400] (Fig. 1).

The NDiV genome organization most closely resembles that of nidoviruses, the only group of ssRNA+ viruses that includes representatives with genomes larger than that of NDiV. This putative relationship was subsequently verified in experimental and bioinformatics analyses of the function and expression of the 3′-ORFs region and in bioinformatics analyses of ORF1a and ORF1b, as described below. The latter studies also provided insights into the evolution and molecular biology of other nidoviruses.

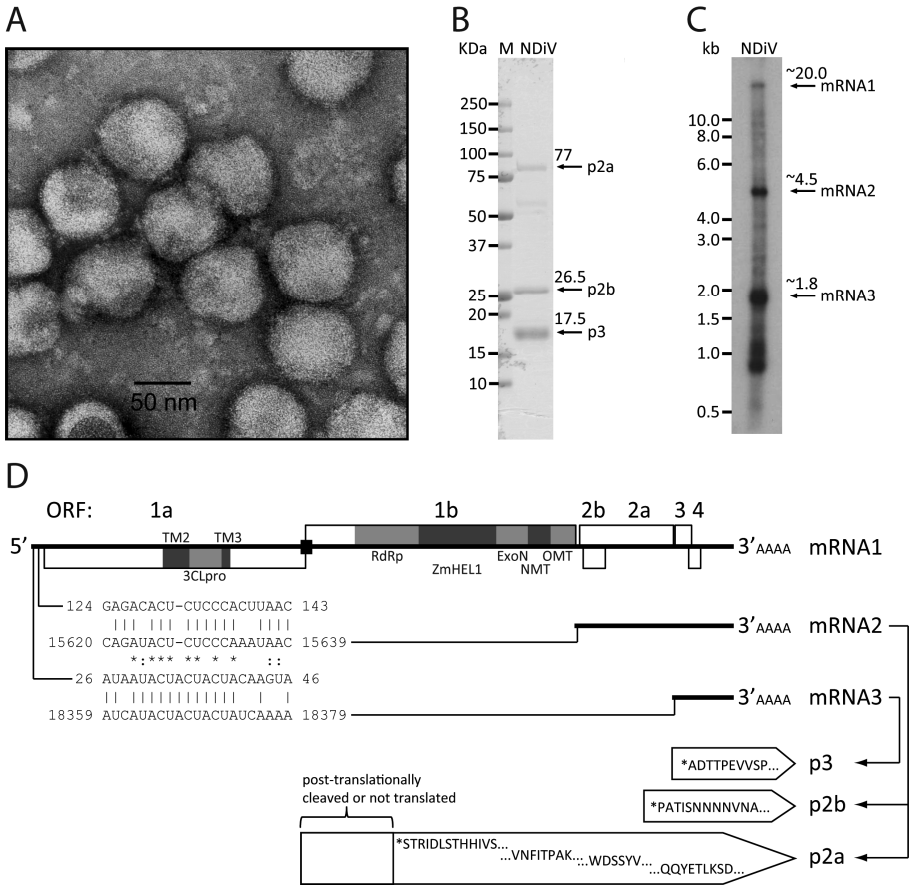


Figure 3. NDiV characteristics. (A) Electron micrograph of negatively stained NDiV virions. (B) SDS-PAGE analysis of NDiV virion proteins. (C) Detection of NDiV genomic (mRNA1) and subgenomic mRNAs (mRNAs 2 and 3) by Northern blot hybridization analysis of total intracellular RNA from virus-infected cells, using a radiolabeled probe complementary to the 3′end of the NDiV genome. (D) NDiV genome organization and expression. Open reading frames (ORFs) are represented by open rectangles and ORF1a- and ORF1b-encoded protein domains identified by bioinformatics analyses (see Table 2) are highlighted in grey. Peptide sequences of virion proteins were determined as described in the Materials and Methods section and mapped to the products of ORFs 2a, 2b, and 3 (bottom-right). N-terminal protein sequences are indicated by (*), other peptide sequences indicate inner sequences. The actual molecular size of the ORF2a product (approximately 77 KDa in SDS-PAGE in B) is considerably shorter than the calculated size (102 KDa), suggesting that p2a may be post-translationally proteolytically processed or that its translation starts at another AUG codon in the ORF. Two pairs of conserved potential TRSs – for sg mRNAs 2 and 3, respectively - were identified in the NDiV genome and aligned (bottom-left), with each pair consisting of a putative leader TRS in the 5′-UTR and a body TRS in the 3′-proximal region of the genome. Between these TRS pairs, eight and three positions include complete match (*) and nucleotide overlap (:), respectively.

Table 1. Mosquito pools collected in Vietnam from which NDiV was isolated.

pool[a]	location	month	Species	quantity[b]
02VN009	Ha Tay	Mar	*Culex vishnui*	25
02VN018	Quang Binh	Mar	*Culex vishnui*	170
02VN178	Quang Binh	Aug	*Culex tritaeniorhynchus*	102
02VN180	Quang Binh	Aug	*Culex tritaeniorhynchus*	83

[a] 359 mosquito pools were collected between Sep. 2001-Dec.2003. The four pools listed in the table were collected in 2002 and infected also with Banna virus.
[b] number of mosquitoes in each pool.

Function and expression of the 3′-ORFs region. Three virion proteins, p2a, p2b, and p3, were assigned to ORFs 2a, 2b, and 3, respectively, by peptide sequencing analysis (Fig. 3D). No significant similarity was found between these ORFs of NDiV and proteins of other origin in BLAST-mediated searches[11]. The p2b protein is highly hydrophilic and enriched with proline (7.5%) and acidic residues (17.8%), and – relative to other virion proteins – with basic residues (7.9%) making it a potential nucleocapsid (N) protein. The p2a and p3 proteins, and the putative protein encoded in ORF4 (p4) contain, respectively, six, two, and two stretches of hydrophobic residues indicative of transmembrane helices (Fig. S2). These proteins also include, respectively, twelve, two, and three potential N-linked glycosylation signals (NXS/T), and fifteen, six, and four cysteine residues that might form disulfide bridges at locations flanked by hydrophobic regions. These characteristics are typical for glycoproteins of other RNA viruses. Based on size considerations, the largest protein, p2a, might be an equivalent of the spike (S) protein, while p3 and/or p4 might be a smaller glycoprotein and an equivalent of the membrane (M) protein of nidoviruses.

We also asked whether NDiV resembles other nidoviruses in using subgenomic mRNAs for expressing the 3′-end ORFs located downstream of ORF1b. First, we attempted to identify potential TRS motifs in the viral genome sequence, which were expected to reside in the 5′-UTR as well as in the regions immediately upstream of ORF2a, 3, and 4. Although no common repeats larger than six nucleotides were identified in these four areas, we noticed the presence of two pairs of near-perfect repeats: the first pair located in the 5′-UTR (nt 26–40 of the genome) and the region upstream of ORF3 (14 out of 15 residues are identical), and the second pair encompassing nt 125–137 of the 5′-UTR and a sequence immediately upstream of ORF2a/2b (12 out of 13 residues are identical) (Fig. 3D). The two pairs share from ~43 to 52% pair-wise sequence identity in an alignment containing a single gap (Fig. 3D), and no other repeats of comparable or larger size were found in the analyzed areas. The locations and sizes of these repeats suggest they are TRS signals, although no candidate TRS was identified immediately upstream of ORF4; to our knowledge, the use of two alternative leader TRSs has not been observed in other nidoviruses thus far. These observations suggested that NDiV uses at least two subgenomic mRNAs for the expression of the 3′-located ORFs and that these mRNAs have 5′-terminal sequences of different size in common with the viral genome.

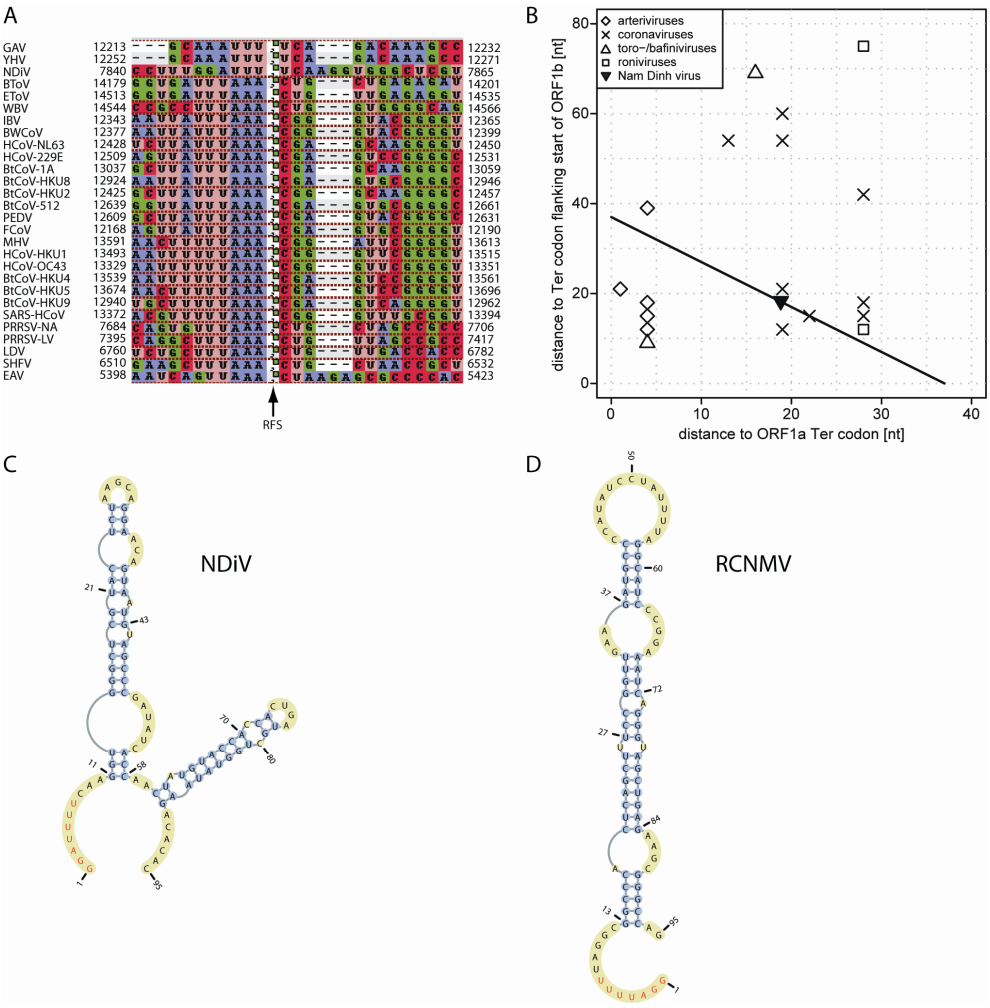


Figure 4. ORF1a/ORF1b ribosomal frameshifting in the NDiV genome. (A) The nucleotide alignment around the open reading frame (ORF) 1a/1b ribosomal −1 frameshift (RFS) in 28 nidovirus species. The alignment column marked with an arrow indicates the RFS position and precedes a nucleotide which is read twice, as the third nucleotide of an ORF1a codon and the first coding nucleotide of ORF1b, respectively. In NDiV this residue is predicted to be located at genome position 7851. Numbers flanking the alignment indicate genomic positions. (B) RFS position in the ORF1a/ORF1b overlap of nidoviruses. The RFS position in selected nidoviruses is plotted as distances between RFS and termination codons flanking ORF1b and ORF1a from upstream and downstream, respectively. In NDiV, the predicted RFS (filled triangle) and all other candidate positions are located on black line in the 40 nt ORF1a/ORF1b overlap. (C and D) RNA secondary structure of sequence fragments around of the ORF1a/ORF1b overlap in NDiV and Red clover necrotic mosaic virus (RCNMV) as predicted by pknotsRG[385]. The putative slippery sequences are in red (see also A). The predicted stem loop in RCNMV closely matches the one presented in [253].

To verify this model, we used a P^{32}-labelled probe complementary to the 3′-end of the NDiV genome in a Northern blot hybridization with total RNA isolated from NDiV–infected C6/36 cells (see Materials and Methods; Table S2, and Fig. 3C). This analysis revealed three prominent RNA species with apparent sizes of about 20, 4.5, and 1.8 kb, which match those expected for the genomic RNA and two subgenomic RNAs, mRNA2 (to express ORF2a and ORF2b) and mRNA3 (for ORF3 and possibly ORF4), respectively. We also observed a set of less abundant bands in the 0.9–1.1-kb size range, whose origin(s) and relevance remain to be established.

ORF1a/ORF1b ribosomal frameshift signal. Nidoviral ORF1a/ORF1b −1 ribosomal frameshifting (RFS) is controlled by a "slippery sequence" and a stem-loop or pseudoknot RNA structure immediately downstream[54]. RFS is conserved in nidoviruses and this property is widely used for computational mapping of its determinants in newly sequenced genomes. We followed this approach to map potential RFS signals in the NDiV genome (Fig. 4). The 40-nt NDiV ORF1a/ORF1b overlap region was found to have the best match (GGAUUUU) with the slippery sequence (AAAUUUU) of invertebrate roniviruses[83], which deviates considerably from the pattern (XXXYYYZ) conserved in vertebrate nidoviruses (Fig. 4A). No appreciated similarity with the latter motif was found in the NDiV ORF1a/ORF1b overlap region. The distances separating the NDiV putative RFS from the termination codons flanking the ORF1a/ORF1b overlap are within the range found in large nidoviruses, while being out of the distance range to the ORF1a stop codon of small nidoviruses (Fig. 4B). According to the analysis of a 190-nt sequence - which starts within the NDiV ORF1a/ORF1b overlap - with Mfold[502] and pknotsRG[385], the predicted slippery sequence is followed by a complex stem-loop structure; no pseudoknots, unless forced, are predicted in this region (Fig. 4C). The slippery sequence, distance to the downstream RNA secondary structure, and predicted fold resemble those of Red clover necrotic mosaic virus (RCNMV), a ssRNA+ plant virus of the family *Tombusviridae*[253,254] (Fig. 4C–D). These results identified the critical elements of the putative NDiV RFS as being most unique among those described for members of the order *Nidovirales*.

Nidovirus-wide conserved domains: TM, 3CLpro, RdRp, Zm-HEL1, and NendoU. Nidoviruses are distinguished from other RNA viruses by a constellation of 7 conserved domains having the order TM2-3CLpro-TM3-RdRp-Zm-HEL1-NendoU, with the first three being encoded in ORF1a and the remaining four in ORF1b. TM2 and TM3 are transmembrane domains, Zm is a Zn-cluster binding domain fused with HEL1, and 3CLpro is a 3C-like protease[174] (however see below). Since NDiV was found to be very distantly related to the other nidoviruses known to date, sequence-based functional characterization presented a considerable technical challenge. In comparative sequence analysis, profile-based methods that employ multiple sequence alignments are known to achieve the best signal-to-noise ratios[11,151,191]. They have been the methods of choice for establishing remote

relations in biology, also in our prior studies of nidoviruses[179,210,229,432]. In this study we used profile vs. sequence and profile vs. profile searches as implemented in HMMer and HHsearch, respectively, for general comparisons. To prepare profiles, we selected representatives of small and large nidoviruses, and also three subsets of large nidoviruses (coronaviruses, toro/bafiniviruses, and roniviruses). Using profile-based searches we identified counterparts (orthologs) of nidovirus-wide conserved enzymatic domains in the NDiV pp1ab. For the identification of TM2 and TM3, predictions of transmembrane helices by TMpred were used.

Six out of the seven nidovirus-wide conserved protein domains, TM2-3CLpro-TM3-RdRp-Zm-HEL1, were mapped in the canonical position and order in the NDiV ORF1a/1b sequence (Table 2). Three of these putative NDiV domains, 3CLpro[13,498], RdRp[445], and HEL1[418] are enzymes conserved in all nidoviruses[169]. They have counterparts of all invariant and highly conserved residues implicated in catalysis in other nidoviruses, a finding indicative of the functionality of these proteins in NDiV.

Like its orthologs in corona- and roniviruses, the NDiV 3CLpro is predicted to employ a catalytic His-Cys dyad. Its substrate-binding site is predicted to include a conserved His residue which was implicated in controlling the P1 specificity for Glu/Gln residues in other viruses, a hallmark of 3C/3CLpros[185]. Surprisingly, despite this finding, no candidate cleavage sites with the characteristic 3CLpro-specific signatures could be identified in the NDiV pp1a/1ab. Consequently, the sizes of all NDiV replicative domains described in this paper (Table 2) are based on the hit sizes in profile searches and are subject to future refinement. Collectively, these results strongly indicate that NDiV encodes all nidovirus-wide conserved replicase domains except for NendoU (Figure 3D; see also below), thus supporting the classification of NDiV as a nidovirus.

Conserved domains common to large nidoviruses: ExoN and OMT. All large nidoviruses express an ExoN[323] of the DEDD superfamily, which is not found in other ssRNA+ viruses, and an OMT[51,96] of the RrmJ family, that is not present in arteriviruses[432]. The presence of these domains therefore discriminates large from small nidoviruses. Using profile searches in the ORF1b-encoded part of pp1ab, homologs of these two enzymes were identified in the NDiV genome (Table 2). Using an ExoN multiple sequence alignment of NDiV and large nidoviruses, the conserved motifs I, II, and III, including the catalytic residues (two Asp and one Glu), as well as the ExoN-specific Zn-finger module were identified in the NDiV ortholog (Fig. 5A). Furthermore, the NDiV ExoN shows an insertion whose size and position correspond to those of the second Zn-finger-like module that is exclusively found in roniviruses. However, unlike the ronivirus domain, NDiV appears to lack His/Cys residues potentially involved in Zn-binding. According to a multiple sequence alignment of nidovirus OMTs (Fig. 5B), the putative NDiV OMT contains motifs X, IV, VI and VIII, encompassing residues of the catalytic KDKE tetrad, as well as motif I involved in

binding of the methyl donor[96]. These data imply that NDiV ORF1b encodes functional ExoN and OMT domains (Fig. 3D), which are both typical of large nidoviruses.

Nidovirus- and large nidovirus-specific domains absent in some lineages: NendoU and NMT. NDiV ORF1b includes a ~750-nt region that is flanked by the upstream ExoN and downstream OMT domains and was expected to encode a NendoU domain[41,232,333,387], given its presence at this locus in all nidoviruses known so far[414,432]. Surprisingly, however, profile searches of nidovirus NendoUs revealed no significant hits in the corresponding region of the NDiV sequence (E-values>9.5). This observation prompted us to re-examine the NendoU assignment in other nidoviruses, including the invertebrate roniviruses[432]. Using profile-sequence and profile-profile comparisons mediated by HMMer and HHsearch, respectively, NendoU counterparts were readily identified in all corona-, toro/bafini-, and arteriviruses (E-values<10^{-4}), but not in roniviruses (E-values>4.5). We therefore conclude that, unlike other (vertebrate) nidoviruses, the invertebrate NDiV and roniviruses do not encode a NendoU domain (Fig. 3D).

We proceeded to analyze this genomic region flanked by ExoN and OMT in invertebrate nidoviruses in more detail. First, using a ronivirus profile vs. NDiV pp1ab sequence comparison, we found that these domains are moderately similar to each other (E-value = 0.18), suggesting a weak conservation of a common function in these newly recognized orthologous domains of NDiV and roniviruses. Their alignment was converted into a profile with which we screened all domains of our in-house nidovirus profile database (see Materials and Methods). Remarkably, the only significant hit (E-value<10^{-4}) was recorded against the coronavirus NMT profile (Table 2). For comparison, its similarities with NendoU profiles of corona-, toro/bafini- or arteriviruses were not significant (E-value>1.5). These data indicate that NDiV and roniviruses may encode an NMT domain that is flanked by ExoN and OMT.

The coronavirus NMT domain was originally mapped to the C-terminal half of nsp14[51,73]. The corresponding domain in toro/bafiniviruses has a much smaller size (80 aa vs. 200 aa). According to our analysis, it has no significant similarity with the NMT of coronaviruses, or the newly recognized putative NMT of roniviruses and NDiV. Based on these observations, we generated an alignment of the NMT domains of corona- and roniviruses and NDiV (Fig. 5C) in order to search for remote cellular homologs. The N-terminal part of the nidovirus NMT includes a conserved methyl donor binding site (motif I), according to the prior assignment for coronavirus NMTs. In line with this observation, a weak hit between nidovirus NMTs and a cellular guanine N7-methyltransferase involving the motif I region was detected in this study. In their C-terminal part, nidovirus NMTs uniquely include four conserved Cys/His residues indicative of a Zn-binding site that may be part of a separate domain (Fig. 5C).

Table 2. Mapping replicative protein domains on the NDiV genome.

nsp[a,b] homolog	start in genome	end in genome	length [aa]	name	description	basis for domain assignment: target[d]	query diversity[e]	method	support[f]
4	3766	4191	142[c]	TM2	transmembrane domain	pp1a	-	TMpred	>500
5	4549	5352	268	3CLpro	3C-like chymotrypsin-like protease[g]	nsp5[g]	roni	HMMer[g]	4e-05[g]
6	5575	5706	44[c]	TM3	transmembrane domain	pp1a	-	TMpred	>500
12	9378	11048	557	RdRp	RNA-dependent RNA-polymerase	pp1b	nido	HMMer	1e-11
13	12177	13388	404	ZmHel1	Zn module+Superfamily 1 helicase	pp1b	nido	HMMer	7e-12
14	13413	14210	266	ExoN	exoribonuclease	pp1b	corona+toro+roni	HMMer	8e-05
14	14211	14912	233	NMT	N7-methyltransferase[h]	nsp14[h]	corona	HHsearch[h]	6e-05[h]
16	14913	15635	242	OMT	2'-O-methyltransferase	pp1b	corona+toro+roni	HMMer	2e-02

[a] Open reading frame (ORF) 1a and ORF1b nucleotide sequences in NDiV were *in silico* translated to obtain the encoded polyprotein (pp) sequences 1a and pp1ab. To map domains in these polyproteins, we employed *HMMer, HHsearch*. The obtained significant hits were mapped back to the genome.

[b] The proteolytic cleavage sites in the NDiV pp1a/pp1ab remain to be identified. To provisionally assign the mapped domains to mature proteins, the names of non-structural proteins (nsp) in SARS-CoV that are autoproteolytically released from pp1a/pp1ab are shown. Note that all replicative domains mapped in NDiV are located in the canonical positions.

[c] Sizes of TM2 and TM3 as determined by *TMpred* may correspond only to small portions of the respective nsp4 and nsp6 proteins.

[d] Portions of pp1ab that were submitted as targets to profiles searches

[e] Shown is the virus diversity range of a query domain profile: the subfamilies *Coronavirinae* (corona) and *Torovirinae* (toro), the family *Roniviridae* (roni) and the order *Nidovirales* (nido). The used profiles for 3CLpro, RdRp, ZmHel1, ExoN, NMT and OMT are part of an in-house nidovirus domain profile database. No query ("-") was used for analyses mediated by *TMpred*.

[f] E-value for *HMMer/HHsearch* based on a database size of 12000 according to the size of the Pfam (version 24.0, October 2009); *TMpred* score otherwise.

[g] The assignment was done according to a profile-vs-profile search in local mode of the domain flanked by TM2 and TM3 in NDiV against the respective ronivirus profile

[h] The assignment was done according to a profile-vs-profile search in global mode of the domain flanked by ExoN and OMT in NDiV and roniviruses against the respective coronavirus.

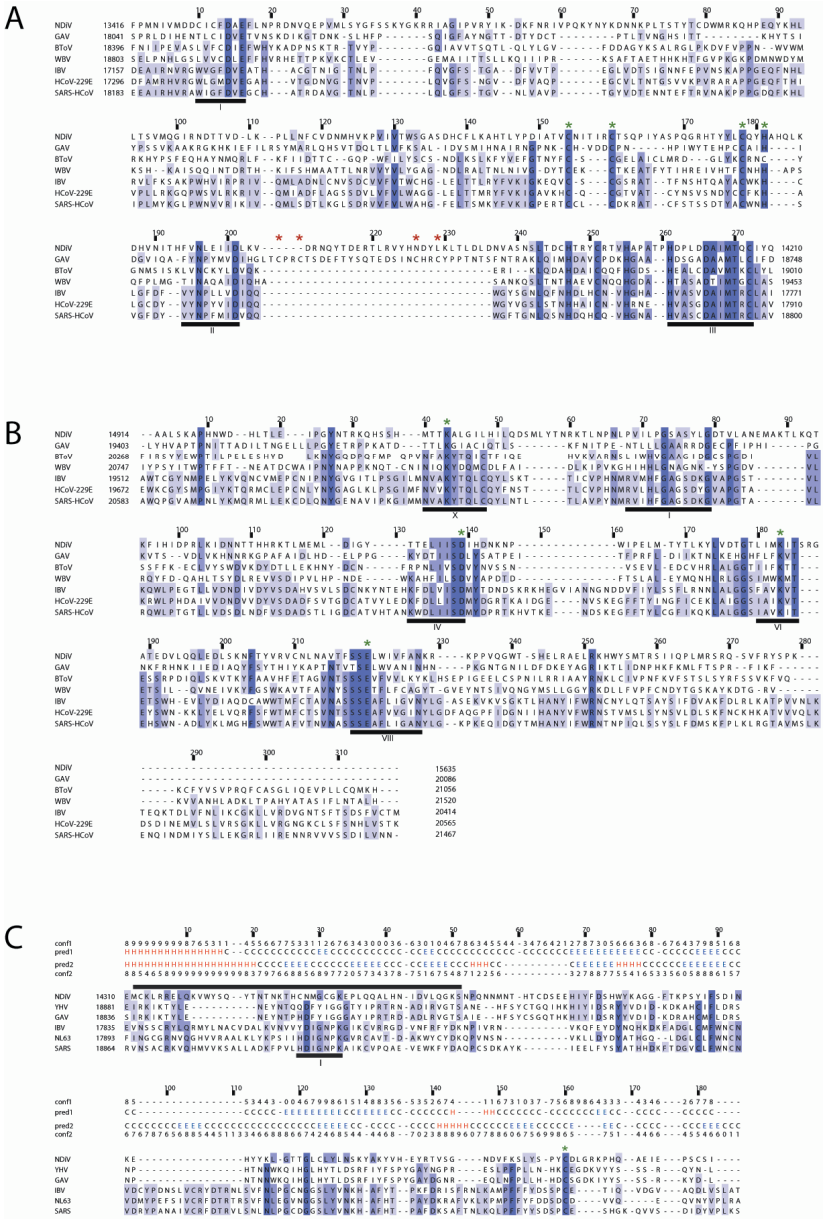

A
B
C
NDiV
GAV
BToV
WBV
IBV
HCoV-229E
SARS-HCoV
YHV
NL63
SARS
conf1
pred1
pred2
conf2

Figure 5. Alignments of ExoN, OMT and NMT domains of NDiV and other nidoviruses. Alignments were compiled utilizing the Muscle program followed by manual inspection. Pictures by JalView[474] and residues are colored according to degree of conservation. Numbers above a column indicate its absolute position in the alignment (start = 1); numbers to the left and to the right of the alignment represent positions in the genome. Selected conserved sequence motifs are highlighted with black bars and roman numbers. (A) In the exonuclease (ExoN) alignment, three motifs are part of the catalytic centre; the domain includes two putative zinc fingers, specific either for roniviruses or for all nidoviruses and highlighted by, respectively, red and green asterisks. (B) In the 2′-O-methyltransferase (OMT) alignment, motifs X, IV, VI and VIII include residues of the catalytic tetrad (KDKE, marked with green asterisks) and motif I is involved in binding of the methyl donor[96]. (C) Protein secondary structure predictions by Psipred[240] for the profiles of N7-methyltransferase (NMT) from 3 NDiV/roniviruses (pred1) and 17 coronaviruses (pred2) and corresponding confidence values (conf1, conf2) were added above the alignment. Only 3 coronaviruses, representing alpha- (HCoV NL63), beta- (SARS-CoV) and gammacoronaviruses (IBV), are shown that results in several empty alignment columns. The black bar on top is a region including the methyl-donor binding site (motif I, delineated by [73]) that gave a hit with a functionally similar site of a cellular guanine N7-methyltransferase (fungus *Encephalitozoon cuniculi*) upon HHsearch of the SCOP database[328] (data not shown). Green asterisks, conserved Cys/His residues that may form a zinc finger.

Collectively these results established a mosaic domain relationship in the pp1ab area flanked by ExoN and OMT domains for large nidoviruses and NDiV. In this genomic region coronaviruses encode both NMT and NendoU domains, while other viruses encode either NendoU (toro/bafiniviruses) or NMT (roniviruses and NDiV).

Phylogenetic analysis of NDiV and other nidoviruses: challenges and approach. Next, we proceeded to determine the phylogenetic position of NDiV among nidoviruses. The phylogeny was inferred using Bayesian posterior probability trees for a concatenated alignment of three enzymes, 3CLpro, RdRp, and HEL1, that are conserved in all nidoviruses (see Materials and Methods). In line with the current nidovirus taxonomy and genomic data[83,162,170,414], this analysis consistently identified the four known major lineages (arteri-, roni-, corona-, and toro/bafiniviruses), as well as a new one represented by NDiV, as the most deeply rooted branches. Our initial attempts to resolve the relationship among the five lineages produced uncertain results. To address this challenge, we adopted a step-wise approach starting from the analysis of close intra-group relationships in the most abundantly sampled subfamily, *Coronavirinae*, and the family *Arteriviridae*, and finishing with an analysis of the most distant inter-(sub)family relationships between the five major lineages. Prior to the nidovirus-wide phylogenetic analysis, the affinity of arteri-, roni-, and toro/bafiniviruses to the subfamily *Coronavirinae* was evaluated through a profile-based analysis involving conserved domains (see Supplementary Text S1 and Table S1). The obtained results confirmed that the strongest sequence affinity exists between corona- and toro/bafiniviruses, which was evident for the 6 out 8 domains that are conserved between coronaviruses and one or more of the other lineages. The HEL1 was the only domain for which an alternative strongest affinity – between corona- and roniviruses – was documented.

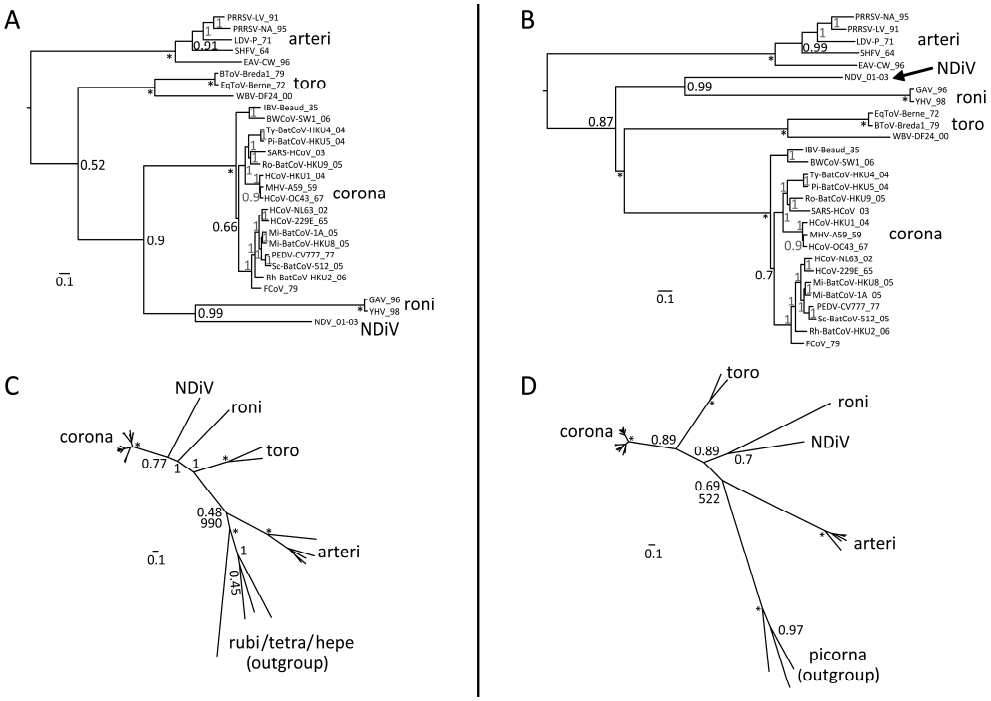


Figure 6. Phylogeny of nidoviruses. To infer phylogenetic relationships between NDiV and other nidoviruses partially constrained trees were calculated using either a concatenated alignment of the three nidovirus-wide conserved domains (A and B) or one nidovirus-wide conserved domain (C and D). For all trees, internal nodes without support value that were fixed prior to the analysis are marked with *, otherwise, numbers indicate posterior probability support values (at the scale from 0 to 1) obtained in either (sub)family- or order-wide analyses (grey and black, respectively). The tree scale bars represent number of substitutions per amino acid position on average. (A) and (B), trees with the constrained topology in which coronaviruses and toro-/bafiniviruses were either fixed as sister clades or not, respectively. Shown are trees for the original alignment which were similar to those obtained for the alignment derivative in which least conserved columns were removed (see Materials and Methods). The trees were rooted on the arterivirus branch. (C) and (D), trees based on conserved HEL1 and RdRp domains, respectively, and including a domain-specific outgroup as described in Materials and Methods. The sister position of coronaviruses and toro-/bafiniviruses was not fixed. For virus listing see trees in A and B. Support values for the outgroup branching in a Maximum-Likelihood analysis with 1000 non-parametric bootstraps, which resulted in an identical topology, is shown below posterior probability support values in both trees. Support values for internal branching within the *Coronavirinae* subfamily and the *Arteriviridae* family are omitted for clarity. The outgroup placement on the arterivirus branch in these analyses was used to root trees in A and B.

Unrooted nidovirus phylogeny. The affinity established above was incorporated as prior knowledge in the nidovirus-wide phylogenetic analysis in order to improve the resolution of the most distant relationships. Accordingly, two alternative reconstructions were conducted with the clustering of toro/bafiniviruses and coronaviruses being either fixed or not. When the clustering was not fixed, roniviruses were found to be closest to coronaviruses (Fig. 6A).

This topology indicated that the HEL1 sequence affinity dominated over that of the RdRp (Table S1) in the concatenated 3CLpro-RdRp-HEL1 alignment. An alternative nidovirus phylogeny was inferred when the clustering of coronaviruses and toro/bafiniviruses was fixed prior to the inference (Fig. 6B). Importantly, in both trees, NDiV was consistently albeit relatively distantly clustered with roniviruses, indicating that this grouping does not depend on the choice of tree-building parameters and is likely genuine.

Rooting of nidovirus phylogeny. To infer the direction of nidovirus evolution, we sought to root the nidovirus phylogeny using an outgroup approach. Neither other viruses nor cellular organisms encode the domain constellation that is conserved in nidoviruses, precluding an expansion of the original nidovirus dataset with outgroup sequences to root the tree. This prompted us to split the domain constellation and perform separate analyses of the evolution of the two most conserved nidovirus protein domains, RdRp and HEL1, which are also among the most conserved in ssRNA+ viruses (Fig. 6C–D). Prior to the analysis, major clades comprising coronaviruses, toro-/bafiniviruses, roniviruses, and arteriviruses, and an outgroup were each fixed to be monophyletic.

For the HEL1 tree (Fig. 6C), the part of the alignment covering the most conserved region from motif I to motif VI (see [175]) was used. Representatives of rubiviruses, betatetraviruses, omegatetraviruses, and hepeviruses were used as an outgroup. The resulting topology closely resembles that of the relaxed nidovirus phylogeny (Fig. 6A), in which vertebrate coronaviruses and invertebrate nidoviruses are sister clades, thus confirming that it is dominated by the HEL1-related component.

For the RdRp tree (Fig. 6D), an alignment of the most conserved RdRp region delimited by motifs G and E (see [184]) was used. Representatives of three divergent picornaviruses (an enterovirus, a parechovirus, and a hepatovirus) were used as an outgroup. The resulting topology matches that of the constrained nidovirus phylogeny (Fig. 6B), in which the grouping of corona- and toro-/bafiniviruses was forced, and could thus be considered RdRp-like.

Despite somewhat incongruent topologies in the two protein-specific phylogenies, in both cases the outgroups are consistently placed at the branch leading to arteriviruses, thus separating small- from large- and intermediate-size viruses in nidovirus evolution. The support for the positioning of the outgroups in the RdRp and HEL1 trees by Bayesian/ML estimates (0.69/522 and 0.48/990, respectively) is relatively low and/or varied in analyses by two methods, possibly due to the very large evolutionary distances separating the major virus groups, including the outgroups. We used the rooting on the arterivirus branch to root the nidovirus tree that was inferred using a concatenated alignment of three domains (Fig. 6A–B).

According to this analysis, small nidoviruses are separated from other nidoviruses, and NDiV is monophyletic with roniviruses in a separate clade of invertebrate nidoviruses, which clusters with the group formed by corona- and toro/bafiniviruses. NDiV and

roniviruses are separated by a large evolutionary distance indicating that NDiV likely is the prototype of a separate family. The topology of the tree in Fig. 6B is compatible with a scenario in which genome size change during nidovirus evolution was dominated by expansion, with contemporary nidoviruses representing different stages in the transition from small to large ssRNA+ genomes.

Discussion

We describe the discovery of an insect-borne ssRNA+ virus, called NDiV, possessing a genome organization, virion properties, mRNAs, and putative proteome characteristics that place it in the order *Nidovirales*. In phylogenetic and protein domain analyses NDiV consistently, albeit relatively distantly, clustered with viruses of the family *Roniviridae*, which seems to make sense biologically given that both infect invertebrate hosts. Although the NDiV classification as the first insect nidovirus is beyond doubt, its characterization was only just initiated in this study. NDiV is likely to possess unique properties concerning, for example, the leader-body junctions of its sg mRNAs and the cleavage sites recognized by its 3CLpro, which both require further characterization.

The principal biological significance of the discovery of NDiV is in the intermediate position this virus occupies between small and large nidoviruses in the genome size distribution observed for ssRNA+ viruses. Prior to this study, the existence of currently circulating nidoviruses with genome sizes within this gap was even highly uncertain (see Introduction). Together small and large nidoviruses cover the upper ~19 kb (~66%) of the entire ssRNA+ genome size range and are separated by ~10 kb (32%). The very existence of NDiV validates the previously established evolutionary relationship between the remotely related arteriviruses and coronaviruses that have very different genome sizes[98]. Characterization of arteri- and coronaviruses by comparative genomics has been instrumental in defining the common and unique features of members of the order *Nidovirales*[169], and has guided the delineation of potential targets for antiviral drug design[183].

The inclusion of NiDV in this analysis yields additional and novel insights with implications for nidoviruses and other RNA viruses at large. It allowed us to revise and expand the assignment for two replicative enzymes of nidoviruses – NendoU and NMT. Prior to this study, the former was considered to be a genetic marker of nidoviruses[432]. Still, its (universal) function in the replication cycle of (vertebrate) nidoviruses has remained enigmatic, despite steady progress in the biochemical, structural, and genetic characterization of this enzyme in arteri- and coronaviruses[40-42,194,232,242,248,333,370,387]. Our analysis showed that invertebrate roniviruses and NDiV do not encode a NendoU domain implying that, contrary to the current paradigm, the utilization of this enzyme in replication may be restricted by the host organism. Surprisingly, and in contrast to the case of NendoU, invertebrate nidoviruses were found to encode a putative NMT, whose ortholog was

previously identified in SARS-CoV and shown to be conserved in the subfamily *Coronavirinae*[51,73]. Our observation indicates that certain aspect(s) of the nidovirus replicative cycle that are controlled by the NMT domain could be similar in coronaviruses and invertebrate nidoviruses, but not toro/bafiniviruses which are otherwise closer to coronaviruses. Collectively, our insights into the phyletic distribution of NendoU and NMT reveal a modularity of some of the major subunits of the replication apparatus in large nidoviruses, which must be rationalized in future mechanistic studies and taken into account in drug development efforts.

Although the NDiV genome size is intermediate between those of small and large nidoviruses, NDiV most closely resembles large nidoviruses in properties that are not universally conserved in the order. Particularly, NDiV does not encode a homolog of the replicative protein of unknown function (nsp12) that is exclusively conserved in arteriviruses[169] and it has a set of three replicative enzymes, OMT, NMT, and ExoN, encoded in large but not in small nidoviruses. These three enzymes are encoded in ORF1b, downstream of the RFS (Fig. 3D and Fig. 4) and in the vicinity of the two key enzymes for RNA synthesis, RdRp and HEL1, with their expression level being downregulated relative to that of the ORF1a-encoded subunits.

Despite these common properties, the two methyltranferases (OMT and NMT) differ from ExoN in their relation to genome size. Particularly, OMTs are known to be also encoded by flaviviruses[128] whose genome size of ~10 kb is average for RNA viruses, while the NMT domain was found to be lacking in a subset of large nidoviruses represented by toro-/bafiniviruses (this study). Furthermore, an N-methyltransferase function, albeit associated with a domain seemingly unrelated to the NMT domain of nidoviruses, was identified in the large Alphavirus-like supergroup of ssRNA+ viruses, whose members have genome sizes from ~7,000 to 19.500 nt[9,321,396]. ssRNA+ viruses use methyltransferases to modify the 5′-end of their mRNAs (cap structure), which was recently found to be essential in the control of translation and innate immunity[89,503]. It is not clear whether the use of methyltransferases may provide particular benefits for genome size control and/or promote genome expansion, although the involvement of OMT in other modifications than 5′-end capping was previously proposed for large nidoviruses[432].

In contrast to the case of the methyltransferases, the link between ExoN and genome size control in nidoviruses is supported by accumulating evidence obtained from different hypothesis-driven genetic studies[99,174]. First, ExoN is exclusively found in a phylogenetically compact cluster of ssRNA+ viruses with large genome sizes. Second, cellular homologs of ExoN control the fidelity of replication in DNA-based life forms and are essential to maintain these large genomes. Third, ExoN active site mutants in MHV and SARS-CoV showed a stable phenotype characterized by a clearly enhanced mutation rate and nearly wild-type progeny yields.

The identification of the ExoN-encoding NDiV further strengthens the case for the direct involvement of ExoN acquisition in genome size expansion. First, because of its

distant relation to any known virus and its insect host range that is a novelty for nidoviruses, NDiV provides an essentially independent verification for the association of ExoN with RNA viruses employing large genomes.

Second, it increases our confidence that no other domain is associated with large genome sizes in nidoviruses as strongly as ExoN is. The existence of such a domain is unlikely but it cannot be formally excluded because the entire proteomes of nidoviruses are yet to be fully described. However, our confidence about the lack of this alternative domain grows with the decrease of difference between genome sizes of nidoviruses containing and lacking ExoN: the smaller this difference the less capacity remains to encode an additional domain. With the identification of NDiV, this genome size gap decreased from ~10.6 kb to ~4.5 kb, the largest drop since this gap could have been recognized (~14.9 kb in 1991) (Fig. S3).

Third, following the discovery of NDiV, only ~0.8 kb remains of the other genome size gap of ~7 kb that previously separated the ExoN-containing nidoviruses from all other ssRNA+ viruses (Fig. 1). Thus, a major step has been made towards a more precise definition of the RNA genome size limit above which the recruitment of a specialized enzyme for replication fidelity control may be a prerequisite. According to a custom binomial test (see Materials and Methods), the probability to observe the association of ExoN and large ssRNA+ genome size by chance may be 10−6 or lower. The genome size threshold of ~20 kb, as defined by NDiV and a closterovirus[106], which has the largest genome size among ssRNA+ viruses other than nidoviruses, is also valid for unsegmented RNA viruses of other classes, all of which do not employ an ExoN in their replicative machinery[217].

The fixation of the ExoN domain in nidovirus genomes may be rationalized in the framework of a unidirectional triangular relationship that includes complexity, replication fidelity (mutation rate), and genome size[131] (Fig. 7). In RNA viruses, the low fidelity of replication severely restricts the size of their genomes, which can encode only relatively simple replication complexes that, hence, suffice to support low-fidelity replication[33,217]. This low-state trap is known as the “Eigen paradox”. Accordingly, a transition from the “low” to the “high” state may not be accomplished by changing only one element of the triangle, e.g. improving replication fidelity, since such a change would not be compatible with the “low” state of the other two elements[131,218]. The exclusive presence of ExoN in ssRNA+ viruses above 20 kb supports the logic of the Eigen paradox[131]. It also shows how the paradox could be solved with a single evolutionary advancement, the acquisition of ExoN, which may have relieved the constraints on all three elements of the triangular relationship (Fig. 7), providing a lasting benefit to the virus lineage that acquired ExoN. This advancement may have been accompanied by an immediate fitness gain. Accordingly, the ExoN acquisition could have provided the ancestral virus with improved control over the fidelity of its replication and the mutation spectrum (quasispecies structure) of its progeny[107,133], which may have facilitated virus adaptation to the environment[99,285]. Alternatively, ExoN could have been acquired in an evolutionarily neutral event. Through subsequent mutation this enzyme might have gained

beneficial properties for the ancestral virus and its progeny. The functional and structural characterization of known nidoviruses and yet-to-be identified viruses in the genome size range around that of NDiV will be required to clarify this key aspect in the transition from small to large nidoviruses.

The acquisition of ExoN by an ancestral nidovirus must have produced viable progeny but it remains unknown whether, besides ExoN, any additional properties of the ancestral nidovirus were critical for genome expansion, as was speculated elsewhere[432]. Recently an exoribonuclease was identified in the ssRNA- arenaviruses, which have genome sizes below 10 kb[202,376]. Unlike nidoviruses, arenaviruses employ the exoribonuclease as a domain of their nucleocapsid protein that, accordingly, mediates a non-replicative function. In line with these differences, the nidovirus ExoN and the arenavirus exoribonuclease do not share specific sequence affinity (CL and AEG, unpublished data), indicating that both are likely to have been acquired from independent sources and were integrated into different genetic settings to perform different functions.

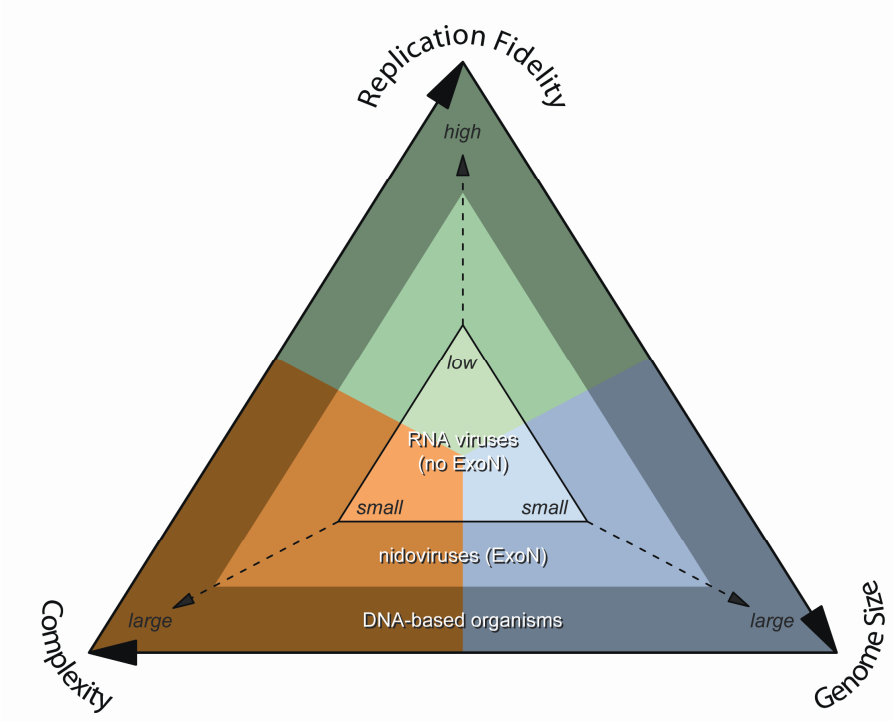


Figure 7. The Eigen trap and a model of nidoviral escape by ExoN acquisition. The scheme depicts the unidirectional relationship between replication fidelity, genome size, and complexity. The vector of variation for the dimensions defined by the three elements of the relationship is shown in a simplified form. The position of RNA viruses in the inner triangular space (Eigen trap) and the proposed effect of ExoN acquisition in nidoviruses on this position are indicated. This is a color version of the original Figure 7 of this publication; DOI:10.1371/journal.ppat.1002215.g007.

NDiV may be the first but likely not the last nidovirus identified in mosquitoes[243]. Systematic probing of these and other insects could lead to the discovery of new nidoviruses, and characterization of those with genomes in the size range between small and large nidoviruses could be particularly insightful. As presented in this study, benefits of these advancements could be multifold and provide a foundation for both fundamental and applied research on newly discovered and already known viruses.

Materials and Methods

Mosquitoes handling for virus isolation. During continued surveillance for JEV in Vietnam between September 2001 and December 2003, 24,097 female mosquitoes belonging to six different *Culex* species (*Culex tritaeniorhynchus*, *Culex gelidus*, *Culex vishnui*, *Culex fusco*, *Culex pseudo*, and *Culex quinquefaciatus*) were collected. They were divided into 359 pools, each containing a single mosquito species and handled with utmost care following the appropriate biosafety measures. For the digestion of blood meals, the samples were kept in 5% glucose for two weeks at room temperature and a humidity of ~90%. The most abundant species was *Culex tritaeniorhynchus* (10,194 mosquitoes accounting for a 42.3% share), followed by *Culex gelidus* (6,199, 25.7%), *Culex vishnui* (3,780, 15.7%), *Culex quinquefaciatus* (2868, 11.9%), with the remaining species ranging from 0.3%–4.1%. Mosquito pools were stored at −70 C prior to processing for virus isolation.

Virus propagation in cell cultures. Four cell lines were used to isolate viruses, but NDiV was evident only in samples from *Aedes albopictus* C6/36 cells grown at 28 C in Eagle's Minimum Essential Medium (EMEM) containing 10% fetal calf serum (FCS) and 0.2 mM non-essential amino acids[228]. Pooled mosquitoes were washed three times in sterile phosphate-buffered saline (PBS, pH 7.2) containing 1000 g/ml each of penicillin and streptomycin, followed by rinsing with antibiotics-free PBS. The homogenates were prepared by triturating the mosquitoes in 2%-FCS-EMEM with subsequent centrifugation at 2,000 g for 10 min. The suspensions were filtered (0.22 nm Millipore, USA) and applied to C6/36 cells, which were monitored daily for cytopathic effects, also after three blind passages. The cell death, probably due to apoptosis, was indeed observed upon NDiV infection. The ICF were clarified by centrifugation at 2,000 g for 10 min.

Genome cloning and sequencing. The nucleic acid was extracted from the purified NDiV virus particles using phenol-chloroform extraction. It migrated as a single band in agarose gel electrophoresis, which was sensitive to RNase but not DNAse treatment, indicative of an RNA virus genome. Accordingly, reverse transcriptase (RT) was used to amplify parts of the NDiV genome by Random Arbitrary Primers-PCR (RAP-PCR) in order to initiate sequence analysis. Cassette primers (C1 and C2) coupled to random hexamers (Hx) were employed.

Following synthesis of first and second cDNA strands with C1Hx and C2Hx primers, respectively, PCR amplification was performed using the cassette primers C1 and C2 as per the standard protocol[475]. Three amplicons of different sizes, which were specific for the virus-containing samples, were then cloned in the pCR2.1-TOPO vector (TOPO TA Cloning Kit, Invitrogen) according to the manufacturer's instructions. The sequence of the first cloned fragment (referred to as “index clone”) was determined by Big Dye Terminator Cycle Sequencing using M13 forward and reverse primers in an ABI 310 or 3100 automated DNA sequencer (Applied Biosystems). The cloned region of the genome was extended by ‘gene walking’ using primers based on previously obtained sequence information (Table S2).

To sequence the genomic region upstream of the index clone, the following amplification strategy was used, involving two DNA fragments called double-stranded (ds) cDNA and anchor DNA. To produce ds cDNA, viral genomic RNA was mixed with 10 mM dNTP mix and 2 pmol of 15-mer gene-specific primers (NDiV-RACE492-477RP, NDiV-RACE302-288RPB and NDiV-RACE435-420RPC) (Fig. S1A, Table S2).

An anchor DNA was synthesized by PCR that amplified a specific fragment of pUC19, including its multiple cloning site (Fig. S1B). Both, the ds viral cDNA and PCR product obtained from pUC19 (anchor) were digested by several restriction enzymes whose sites are present in the pUC19 multiple cloning site (BamHI, EcoRI, KpnI, HindIII, ScaI, and PstI). The digested pUC19 PCR products were then purified using the QIAXII gel purification kit (Qiagen) in order to collect the longer DNA fragments. The digested viral cDNAs were also purified by filtration using Micropure-EZ (Millipore) and Microcon YM-100 (Millipore) to remove enzymes and buffers. In a next step, the purified cDNAs and anchor DNAs were mixed and ligated using T4 DNA Ligase (TaKaRa). The unknown region of viral cDNA was then amplified by semi-nested PCR using LA-taq (TaKaRa), two viral gene specific primers and one pUC19 primer (Table S2) as shown in Fig. S1C. The reaction process included an initial denaturation at 96°C for 5 min, 35 cycles at 96°C for 30 sec, 53°C for 30 sec, and 72°C for 7 min, and a final extension at 72°C for 10 min.

The known viral genome sequence was further extended by long RT-PCR which resulted in an 8 kb fragment with a 68-nucleotide polyA tail representing the 3′-end of the NDiV genome. The GeneRacer™ Kit (Invitrogen) was used to sequence the 5′-end of the NDiV's genome.

The NDiV origin of newly obtained sequences was further validated by probing different samples with a primer pair designed against the index clone. This pair of primers recognized NDiV isolates, but not JE and dengue viruses (flaviviruses) or SARS-coronavirus (Coronavirus). These results indicated that NDiV is a novel mosquito virus.

RNA probe generation for Northern blotting analysis. Specific primers encompassing NDiV nts 19,733 and 20,126 (including 2 Adenines of the poly (A) tail), respectively, were designed (Table S2). The generated PCR product was purified using the Qiaex II gel extraction kit (500) (Qiagen) following the manufacturer's instructions. The purified PCR

product was then ligated to a 3.5 kb plasmid (PCR-XL-TOPO) using the TOPO XL PCR cloning kit (Invitrogen, applying the TA rule based on the Taq polymerase's capacity of adding an extra A at the 3′ end of each DNA chain of a PCR product) as per the manufacturer's indications. Heat shock transformation into One Shot Top 10 chemically competent cells (Invitrogen) was carried out and the transformed cells were incubated in SOC medium at 37 C for 2 hrs. After that, the E. coli cells were cultured in 50 µg/ml containing LB plates overnight and the positive clones were subsequently cultured in LB broth at 37 C overnight. The plasmid alkaline extraction was done using the QIAprep spin Miniprep kit (Qiagen) as the manufacturer indicated. As a next step, verification of the probe orientation was carried out by nucleotide sequencing. Finally, transcription of the cloned DNA sequences was done to generate the RNA probe (in both sense and reverse orientations). The RNA probe was then labeled with ^{32}P by using the AmpliScribe T7 High Yield Transcription Kit (EPICENTRE Biotechnologies) following the company's instructions.

Northern blotting. To investigate the possibility that NDiV generates set of 3′-coterminal sub-genomic mRNA's during its replication, *Aedes albopictus* C6/36 cells were infected with NDiV. Three to four days after infection intracellular poly (A)-containing RNA from mock-infected and NDiV-infected cells was prepared using Dynabeads oligo$(dT)_{25}$ (Dynal Biotech) as per the manufacturer's instructions. RNA was separated on a glyoxal-based agarose gel system and blotted on a positively charged nylon membrane (BrightStar-Plus membrane). The mRNA bands were then hybridized with an α-^{32}P-multiprime-labeled RNA probe specific for NDiV at 65°C overnight (see above RNA probe generation). The membrane was then washed with low and high stringency wash solutions and the RNAs were analyzed by autoradiography. All reagents for mRNA separation, transfer and hybridization (with the exception of the RNA probe) were provided with the NorthernMax-Gly Kit (Ambion). The manufacturer's instructions were followed. A 0.5–10 Kb RNA Ladder (Invitrogen) was used as a marker set to calculate apparent molecular mass of the analyzed bands.

Electron microscopy of virions. For electron microscopy, virus was concentrated from ICF by centrifugation at 12,000 g for 30 min at 4 C, after which 6.6% polyethylene glycol 6000 and 2.2% NaCl were added to the supernatant. After stirring for 1 h at 4 C and centrifugation at 12,000 g for 1 h, the supernatant was discarded. The virus-containing pellet was dissolved in saline-Tris-EDTA buffer, sedimented at 250,000 g for 1 h and resuspended a second time. The concentrated virus was negatively stained with 1% sodium phosphotungstic acid, pH 6.0, and examined at 100 KV using a transmission electron microscope (JEM-100CX, JEOL, Japan)[203].

Sequencing of virion peptides. Virions were purified in a 15–50% sucrose density gradient using an SW32Ti rotor (Beckman Coulter, Inc., Fullerton,CA) at 20,000 rpm for 12–16 h at 4°C. Gradient fractions were analyzed by 16% SDS-polyacrylamide gel electrophoresis and

Coomassie Brilliant Blue G staining (Fig. 2B). Protein bands were excised and either directly sequenced by automated Edman degradation (Applied Biosystems model 491cLC) or digested with lysylendopeptidase prior to HPLC purification and sequencing.

Bioinformatics databases. Genome sizes of ssRNA+ viruses were retrieved from the NCBI Viral Genome Resource[25]. GenBank, version 178.0[37], Pfam database, version 24.0[151], SCOP70, version 1.75[328], and an in-house nidovirus domain profile database[183,432] updated in this study were used to identify putative functional domains encoded by the NDiV genome. Representatives of the nidovirus species defined according to (http://www.ictvonline.org/virusTaxonomy.asp?version=2009) plus NDiV, whose taxonomical status remains provisional, were used as detailed in Table S3. Species names of coronaviruses were taken from ICTV proposal 2008.085-122V.U that was approved by ICTV in 2009. Fields after the "_" sign in virus abbreviations represents sampling year or period.

Basic bioinformatics analyses. The NDiV ORFs were compared with sequence databases using psi-BLAST[11], HMMer 2.3.2[125], TMpred[213], or HHsearch[434]. Protein secondary structure predicted by Psipred[240] was included in the HHsearch-mediated profile searches. RNA secondary structure analysis was conducted using Mfold[502] and pknotsRG[385]. MUSCLE[126] was used to produce alignments of nidovirus proteins that were manually refined in poorly conserved regions. Alignment derivatives, with the least conserved columns removed[65], were prepared using BAGG[18] and were used for profile searches and phylogenetic analyses. Alignments were prepared for publication using JalView[474]. To compile and plot most graphs and conduct statistical analyses we used the R package[377].

Identification of TRS candidates. Using the de novo repeat detection program RepeatScout[372] a library of perfect repeats with unit sizes ranging from four to the maximum observed size of 16 was compiled for the NDiV genome sequence. The library was filtered to retain repeats of different types according to the following constraints applied to each type separately: (i) one repeat copy must be located upstream of ORF1a, and (ii) another one must reside within the 300 nt region immediately upstream of either ORF2a, ORF3, or ORF4. Each set of the retrieved repeats was subsequently analyzed for conservation by alignment that included flanking regions of 20 nt at each side. The longest repeats with highest similarity were considered TRS candidates.

Profile-based similarity searches. To map major nidovirus replicative proteins to pp1ab of NDiV we applied alignment-based methods. Multiple sequence alignments represent a general tool to infer both common ancestry (orthology) of residues for several related sequences (these residues form a fully occupied alignment column) and identify insertion/deletion events (corresponding to alignment columns containing gaps in selected sequences). Multiple alignments can be converted into profiles, which are statistical models

that capture the degree of conservation and the likelihood to observe a certain residue or gap in each alignment column. One type of profiles are profile Hidden Markov Models (HMMs)[273] that are particularly suitable for searching for remotely related sequences (like NDiV which presumably represents a new virus family) in a probabilistic framework. They are implemented, for example, in the programs HMMer and HHsearch which were utilized in this study. A profile HMM can be compared to other HMMs or used to search for motifs in a single sequence. Due to the high degree of divergence of nidovirus sequences, we used alignments of amino acid sequences and profiles derived from these alignments to probe relation between proteins in this study.

Phylogenetic analyses. Phylogenetic analyses were performed as described previously[490]. Bayesian posterior probability trees were compiled utilizing BEAST[119] under the WAG amino acid substitution matrix[478] using Tracer[120] to verify convergence. For the nidovirus-wide analysis, whose sampling is detailed Table S3, we used a concatenated alignment of 3CLpro, RdRp, and HEL1 including 910 aa positions and its derivative of 604 aa positions, from which least conserved columns were removed. In this analysis, the uncorrelated relaxed molecular clock approach (lognormal distribution)[118] was used as it was favored[165] over the strict molecular clock (log10 Bayes factor of 13.6) and equal to the relaxed molecular clock approach with exponential distribution (log10 Bayes Factor of 0.0). Selected internal nodes were fixed using results of separate analyses of subsets of nidoviruses. For phylogenetic analysis of the subfamily *Coronavirinae* and the family *Arteriviridae*, we used respective datasets incorporating between one and three sequences per species and including concatenated alignments of ORF1ab domains that are conserved in each of these groups. The datasets included 35 and 10 sequences for corona- and arteriviruses and consisted of 2302- and 2882-aa alignment positions, respectively. The topologies of these trees closely follow those published[170]. They were used to fix internal nodes in corona- and arterivirus clusters in the subsequent nidovirus-wide phylogenetic analysis. The exception was the basal nodes corresponding to the grouping of the *Alpha*-, *Beta*-, and *Gammacoronavirus* genera and the root of arteriviruses (EAV or SHFV), which were left unfixed. Maximum Likelihood trees were compiled utilizing the PhyML software[196]. The WAG amino acid substitution matrix and rate heterogeneity among sites (8 categories) were applied and support values for internal nodes were obtained using the non-parametric bootstrap method with 1000 replicates. Trees were rooted using domain-specific outgroups: for RdRp, three picornavirus representatives (accession numbers: NC_001489, NC_001897, NC_002058); for HEL1, four rubi-/ tetra-/ hepevirus representatives (NC_001545, NC_001990, NC_005898, NC_001434).

Association of ExoN and large genome sizes. We sought to statistically define a genome size threshold that separates ExoN-containing from ExoN-lacking ssRNA+ viruses. To this end, we developed a custom test employing the binomial probability function and including

all 43 virus groups displayed in Fig. 1. These groups consist of thousands of viruses that are believed to have emerged from a common ancestor, implying that they are not independent. Their dependence varies in virus pairs but, generally, for each virus pair is inversely proportional to the pair-wise evolutionary distance. To account for the dependence of these sequences in our test is technically challenging. To circumvent this problem, we have created a derivative of the virus dataset in which each virus family/group is represented by a single virus, in total 43 viruses. We considered the sequences of these representatives to be essentially independent due to the (extremely) large divergence that is observed, even in the most conserved genes (e.g. see Fig. 6), the lack of recognizable similarity in other genes, and the accompanied gene loss and gain.

For a given genome size threshold, ssRNA+ viruses were partitioned into two groups (below and above that threshold) and the value of the binomial density function was calculated for both groups using information on the presence or absence of ExoN. The final probability of the test is the product of the binomial probabilities for the two groups. We used a binomial success probability of 4/43 since four out of the 43 ssRNA+ virus lineages (NDiV, toro-/bafiniviruses, coronaviruses, and roniviruses) employ ExoN. The test was applied to each possible threshold separating two unique ssRNA+ genome sizes, in total – 42 thresholds. The threshold of ~20 kb, between the genome sizes of NDiV and closteroviruses, gave the lowest probability to observe the ExoN association by chance. We consider the obtained value (10^{-6}) as an underestimate of the true probability that should be calculated by taking into account the sequence dependence and all viruses in the 43 groups, which without exception conform to the ExoN distribution observed in the selected virus representatives used now.

Accession numbers. RefSeq accession numbers of proteins referred to in the text for a selection of prototype nidoviruses are: 3C-like proteinase (EAV: NP_705584, SARS-CoV: NP_828863, WBV: YP_803213, GAV: YP_001661453), RNA-dependent RNA polymerase (EAV: NP_705590, SARS-CoV: NP_828869, WBV: YP_803213, GAV: YP_001661452), superfamily 1 helicase (EAV: NP_705591, SARS-CoV: NP_828870, WBV: YP_803213, GAV: YP_001661452), exoribonuclease (SARS-CoV: NP_828871, WBV: YP_803213), N7-methyltransferase (SARS-CoV: NP_828871), uridylate-specific endonuclease (EAV: NP_705592, SARS-CoV: NP_828872, WBV: YP_803213) and 2′-O-methyltransferase (SARS-CoV: NP_828873, WBV: YP_803213).

Acknowledgments

We thank Ellie Ehrenfeld for the critical reading of the manuscript, Igor Sidorov for discussions, Corrine Beugeling and Miki Higashi for help with genome and peptide sequencing, respectively, Alexander Kravchenko and Dmitry Samborskiy for Viralis management. Also constructive criticisms of three anonymous reviewers, which helped us to improve the manuscript, are gratefully acknowledged. This work was partially supported by Program of Founding Research Centers for Emerging and Reemerging Infectious Diseases, MEXT-Japan (to K.M.), the Netherlands Bioinformatics Center BioRange Program and Leiden University Fund (to A.E.G.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: KM. Performed the experiments: PTN MdCP MP TN FY NTT SI TI KO AI. Analyzed the data: MdCP EJS KM AEG. Contributed reagents/materials/analysis tools: AEG. Wrote the paper: CL MdCP KM EJS AEG. Conceived and designed computational research: AEG CL. Conducted computational research: CL. Analyzed the computational data: CL AEG.

Supporting Information

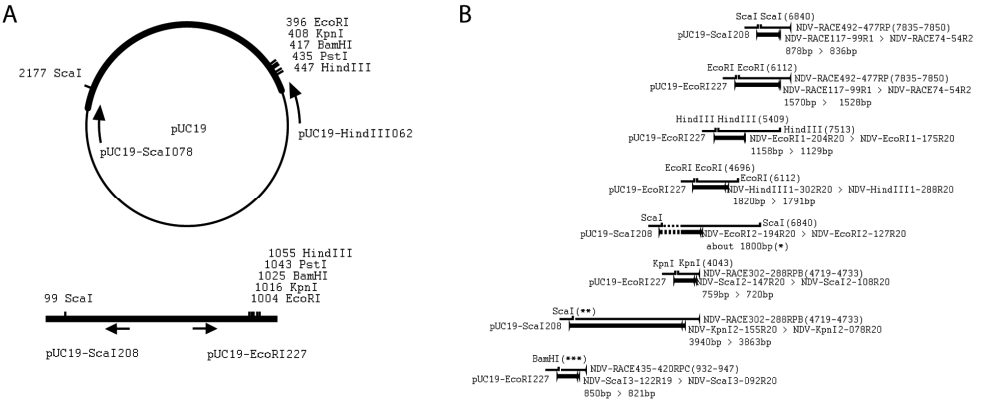


Figure S1. Cloning and sequencing details. (A) To obtain RT-PCR products containing unknown NDiV sequences upstream of the previously sequenced region of the genome the following was done: cDNAs of the NDiV RNA were converted into ds cDNAs, which were digested by restriction enzymes and subsequently ligated to an anchor DNA using those existing restriction sites. For a detailed explanation of each procedure please read the "Genome cloning and sequencing" section of Materials and Methods. (B) Semi-nested PCR was conducted for the anchored ds cDNA of NDiV using one pUC19 specific sense primer (primer pUC119-scaI208 was used for the ScaI-digested sample and primer pUC19-EcoRI227 was used for the samples digested with all the other restriction enzymes) and two reverse gene-specific primers (GSPs) of NDiV for each experiment. The PCR products contained the unknown sequence between GSP and anchor. This process was repeated eight times, and this protocol allowed us to read a total of 7164 bp. The name of each restriction enzyme is followed by its position written in brackets as explained below. NDiV-RACE117-99R1>NDiV-RACE74-54R2 means "primer for first PCR>primer for nested PCR". 878 bp>836 bp means "size of the first PCR>size for the nested PCR". If 1, 2, or 3 asterisks are in brackets, non-specific cuts took place with the following details: (*) non-specific cut and ligation occurred at 4433 bp; (**) non-specific anchoring at 513 bp; (***) In the eighth step, anchor DNA of BamHI and HindIII attached to same location, and it was suggested that reverse transcription stopped there. The GeneRacer (TM) Kit (Invitrogen) was used to read the remaining 205 bp toward the 5′-end of genomic RNA.

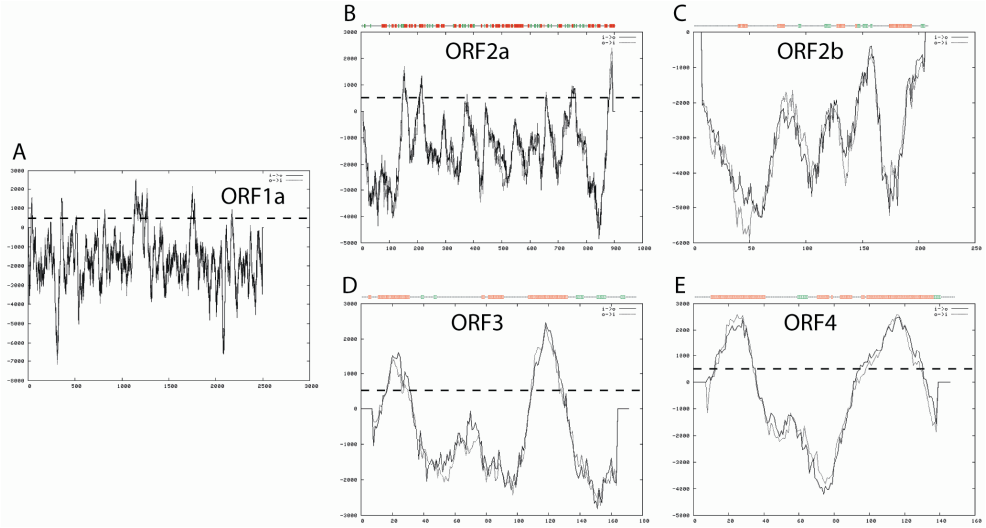


Figure S2. Hydrophobicity plots and secondary structure predictions for (presumed) NDiV structural proteins. Hydrophobicity was calculated using TMpred for the pp1a replicase precursor (ORF1a; A) that served as a control for four (putative) virion proteins p2a (ORF2a; B), p2b (ORF2b; C), p3 (ORF3; D) and p4 (ORF4; E). Horizontal dashed lines depict the threshold (value of 500) for significant association with transmembrane helices. On top of the plots for the structural proteins, Jpred-mediated secondary structure predictions are shown. Predicted alpha helices and beta strands are highlighted in red and green, respectively.

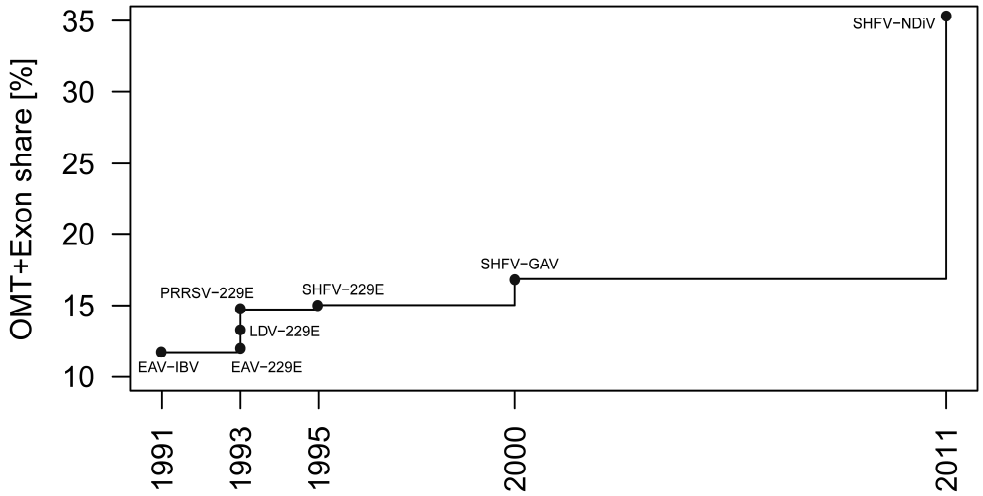


Figure S3. Association of ExoN with large nidovirus genomes and progress in nidovirus genomics. At the X axis, a timeline of nidovirus genome sequencing is plotted. It starts in 1991 when the first pair of genome sequences for both small and large nidoviruses, arterivirus (EAV) and coronavirus (IBV), respectively, became available. The genome size difference (gap) between these viruses was ~14.9 kb. All subsequent time points were selected because new nidoviruses with either larger (for small nidoviruses) or smaller (for large nidoviruses) genomes were released in these years. (For the purpose of this analysis, NDiV was treated as a large nidovirus). As a result, the genome size gap shrunk, in total six times since 1991 (three times in 1993). Currently, the gap that remains is ~4.5 kb (the arterivirus SHFV vs. NDiV). Large nidoviruses are assumed to have acquired a unique genomic region during the expansion of their genome. It includes ExoN and OMT, and some other genes, like the NMT that is found in some large nidoviruses. This region may also include additional domains, due to the thus far incomplete characterization of the nidovirus proteome; they might include one or more with the phyletic distribution characteristic of ExoN and OMT. Understandably, as the genome size gap between large and small nidoviruses has been shrinking due to the discovery of new nidoviruses, the probability that such genes exist is decreasing. Likewise, the share of the size of the ExoN and OMT domains in the total genome size gap could be considered a measure of confidence for the role of these genes in nidovirus genome expansion. At the Y axis, the growth of this share is plotted; it gradually increased from ~12% in 1991 (EAV vs. IBV) to ~35% in 2011 (SHFV vs. NDiV). By far the biggest increase (~17%), and hence the largest gain in support for the role of ExoN in nidovirus genome expansion, was achieved by the sequence analysis of the NDiV genome. The above numbers outline a trend and this analysis should not be confused with a probabilistic framework.

Table S1. Affinity of corona- and toro-/bafiniviruses[a].

coronavirus protein profile[b]	target[c]			
	corona[d]	toro/bafini	roni	arteri
ADRP	e-105	**e-25**	330	34
PL2pro	e-126	*40*	110	160
3CLpro	e-244	*24*	88	43
primase	e-152	*4*	27	200
RdRp	~0	**e-14**	0.004	0.079
HEL1	e-227	e-6	**e-13**	0.81
ExoN	e-244	**0.008**	0.62	38
NMT	e-204	22	*13*	46
NendoU	e-133	*e-4*	23	*e-4*
OMT	e-245	**e-8**	0.29	57
S[e]	~0	**0.005**	100	240
M	e-158	**1**	990	12
E	e-37	56[f]	250	*11*
N	e-247	12000	7000	*2800*

[a] *HMMer* profile searches (global profile against local sequence) were used to determine closest nidovirus relatives of a selection of proteins expressed by coronaviruses. E-values are based on a database size of 12000 according to the size of the Pfam (version 24.0, October 2009). The best hit against the profile of the coronavirus protein alignment is indicated in italic and unique, significant (E-value <= 1) best hits against toro-/bafini- or roniviruses in bold
[b] a profile of an alignment containing 17 coronavirus species was used. ADRP, ADP-ribose-1"-phosphatase; PL2pro, papain-like proteinase 2; 3CLpro, 3C-like proteinase; RdRp, RNA-dependent RNA polymerase; HEL1, superfamily 1 helicase; ExoN, 3'-to-5'exoribonuclease; NMT, N7-methyltransferase; NendoU, uridylate-specific endonuclease; OMT, 2'-*O*-methyltransferase.
[c] numbers represent E-values of a *HMMer* search against the coronavirus protein profile
[d] E-values of hits against coronaviruses itself are shown for comparison
[e] only the C-terminal part of the coronavirus S protein alignment (S2) was used as a profile
[f] hit is against the torovirus M protein

Table S2. Primers used to sequence the 5'-end of the NDiV genome.

Primer	Sequence	Application
NDiV-RACE492-477RP	AAATCCAAAGGGTGCT	cDNA generation
NDiV-RACE117-99R1	GCGTTCAAAATAGCCAAGT	Semi-nested PCR
NDiV-RACE74-54R2	TAGTAATAGCCTTCAGGATCG	Semi-nested PCR
NDiV-EcoRI1-204R20	ATCGAGTGTGTCTGGAGTAG	Semi-nested PCR
NDiV-EcoRI1-175R20	TCAAATAGCGTACGAGTTCA	Semi-nested PCR
NDiV-HindIII1-288R20	AAGATGATGGAGCTAAGGAT	Semi-nested PCR
NDiV-HindIII1-302R20	TGTGGGGGATTGTAAAGATG	Semi-nested PCR
NDiV-EcoRI2-127R20	GGATGTAAGCTGATATGTGG	Semi-nested PCR
NDiV-EcoRI2-194R20	TTTGTTTAGTTCCGTGTCGT	Semi-nested PCR
NDiV-ScaI2-108R20	TGTGAAATTGAGGGGTTTGA	Semi-nested PCR
NDiV-ScaI2-147R20	ATTAGAGGGTTAATGGCAAC	Semi-nested PCR
NDiV-RACE302-288RPB	AGAAGCCCCTTACCA	cDNA generation
NDiV-KpnI2-078R20	CTGGTGCAGACGTACGGAAT	Semi-nested PCR
NDiV-KpnI2-155R20	TTAGGTAGTTTGGTCGTTGT	Semi-nested PCR
NDiV-RACE435-420RPC	TCGCTTACTGCTTTCT	cDNA generation
NDiV-ScaI3-122R19	GAAAAATGTTTAGGCGAGA	Semi-nested PCR
NDiV-ScaI3-092R20	GCAAAAACTGGTGTTTGATA	Semi-nested PCR
pUC19-EcoRI227	ACAGATGCGTAAGGAGAAAA	Semi-nested PCR
pUC19-ScaI208	AACGCTGGTGAAAGTAAAAG	Semi-nested PCR
pUC19-HindIII062	TATGCTTCCGGCTCGTATGT	Anchor
pUC19-ScaI078	AGTAAGTTGGCCGCAGTGT	Anchor
NDiVpolyA-R	TTACCTGTAATGCCAAGCGC	Northern blot*
NDiV19733-F	CGCCTGTAAGAGAGATTGTA	Northern blot*

Table S3. Genome sequences of a representative set of the Nidovirus species.

species name[a]	virus abbreviation[b]	(sub)family	acc. number
Nam Dinh virus	NDiV_01-03	-	-
Gill-associated virus	GAV_96	Ronivirus	AF227196
Yellow head virus	YHV_98	Ronivirus	EU487200
White bream virus	WBV-DF24_00	Torovirus	NC_008516
Equine torovirus	EToV-Berne_72	Torovirus	X52374
Bovine torovirus	BToV-Breda1_79	Torovirus	NC_007447
Human coronavirus 229E	HCoV-229E_65	Coronavirus	NC_002645
Human coronavirus NL63	HCoV-NL63_02	Coronavirus	DQ445911
Miniopterus bat coronavirus 1	Mi-BatCoV-1A_05	Coronavirus	NC_010437
Rhinolophus bat coronavirus HKU2	Rh-BatCoV-HKU2_06	Coronavirus	NC_009988
Miniopterus bat coronavirus HKU8	Mi-BatCoV-HKU8_05	Coronavirus	NC_010438
Scotophilus bat coronavirus 512	Sc-BatCoV-512_05	Coronavirus	DQ648858
Porcine epidemic diarrhoea virus	PEDV-CV777_77	Coronavirus	NC_003436
Geselavirus	FCoV_79	Coronavirus	NC_007025
SARS-related coronavirus	SARS-HCoV_03	Coronavirus	AY345988
Tylonycteris bat coronavirus HKU4	Ty-BatCoV-HKU4_04	Coronavirus	EF065505
Pipistrellus bat coronavirus HKU5	Pi-BatCoV-HKU5_04	Coronavirus	EF065509
Rousettus bat coronavirus HKU9	Ro-BatCoV-HKU9_05	Coronavirus	EF065513
Human coronavirus HKU1	HCoV-HKU1_04	Coronavirus	AY884001
Betacoronavirus 1	HCoV-OC43_67	Coronavirus	AY585228
Murine coronavirus	MHV-A59_59	Coronavirus	AY700211
Avian coronavirus	IBV-Beaud_35	Coronavirus	NC_001451
Beluga whale coronavirus SW1	BWCoV-SW1_06	Coronavirus	EU111742
Equine arteritis virus	EAV-CW_96	Arterivirus	AY349167
Simian hemorrhagic fever virus	SHFV_64	Arterivirus	NC_003092
Lactate dehydrogenase-elevating virus	LDV-P_71	Arterivirus	U15146
Porcine respiratory and reproductive syndrome virus, North American type	PRRSV-NA_95	Arterivirus	AF176348
Porcine respiratory and reproductive syndrome virus, European type	PRRSV-LV_91	Arterivirus	M96262

[a] species names of coronaviruses taken from ICTV proposal 2008.085-122V.U that was approved by ICTV in 2009.
[b] field after the "_" sign represents sampling year or period

Text S1. Sequence similarity-based clustering of corona- and toroviruses.

There is a consensus in the field that coronaviruses and toro/bafiniviruses are nidoviral sister lineages. This relationship has been codified in nidovirus taxonomy with these two groups of viruses forming the two subfamilies in the family *Coronaviridae*. Yet, prior phylogenetic analyses using either RdRp or HEL1 were not as conclusive about this clustering[162], which prompted us to verify it using an alternative approach. We sought to use similarity of domains that are conserved in the subfamily *Coronavirinae* and (partly) shared with other nidoviruses to rank toro/bafiniviruses, roniviruses and arteriviruses in relation to the subfamily *Coronavirinae*. We compiled HMMER profiles for 14 protein domains of 17 coronaviruses, representing replicative proteins (10 domains: ADP-ribose-1''-phosphatase (ADRP), papain-like proteinase 2 (PL2pro), 3CLpro, primase, RdRp, HEL1, ExoN, NMT, NendoU and OMT) and virion proteins (4 domains: S, M, E and N) that together account for ~45% of the ~29kb genome. They were compared in the global profile vs. local sequence mode against products of all ORFs encoded by a representative set of nidoviruses.

The obtained E-values of the top hits for four phylogenetic groups, corona-, toro/bafini-, roni, and arteriviruses were compared for each protein domain (Table S1). Eight domains (6 replicative and 2 virion domains) produced significant hits outside coronaviruses: all 8 with toro/bafiniviruses, and 4 different domains with, separately, roniviruses and arteriviruses. Based on the best hit E-values, toro/bafiniviruses were ranked the top for 6 domains (ADRP, RdRp, ExoN, OMT, S, and M), shared the top spot for one domain with arteriviruses (NendoU) and were ranked second after roniviruses for the HEL1 domain. According to another analysis that is presented in Fig. 5C, corona- and roniviruses but not toroviruses also share an NMT domain. However, this conservation was too remote to be identified by the HMMER-based analysis and it was not included in Table S1. Regardless of considerations involving the NMT domain, the presented results confirm a special sequence affinity between coronaviruses and toro/bafiniviruses among nidoviruses.

CHAPTER 6

The footprint of genome architecture in the largest genome expansion in RNA viruses

Chris Lauber
Jelle J. Goeman
Maria del Carmen Parquet
Phan Thi Nga
Eric J. Snijder
Kouichi Morita
Alexander E. Gorbalenya

manuscript in preparation

Abstract

Small genome sizes of RNA viruses (2 to 32kb) have been linked to the high mutation rate during RNA replication that is thought to lack proof-reading. This paradigm is now being reviewed owing to the discovery of a 3'-to-5'exoribonuclease (ExoN) in nidoviruses, a monophyletic group of viruses with non-segmented, single-stranded RNA genomes of positive polarity and conserved genome architecture. The ExoN, homolog of a canonical DNA proof-reading enzyme, is exclusively encoded by nidoviruses with genomes larger than 20 kb. All other known non-segmented RNA viruses employ smaller genomes. Here we use evolutionary analyses to show that the two- to three-fold expansion of the nidovirus genome was accompanied by a vast amount of replacements in conserved proteins at the scale observed in the Tree of life. To unravel common patterns of such genetically diverse viruses, we exploited functional conservation of the nidovirus genome architecture. This conservation allowed us to partition each genome into five spatially collinear regions in an alignment-free manner. Each genomic region was analyzed for its contribution to genome size change under both linear and non-linear conditions. The non-linear model statistically outperformed the linear one and captured >92% of data variation. Accordingly, individual nidoviruses were found to have reached different points on a common expansion trajectory dominated by three consecutive, region-specific size increases. Our findings indicate a hierarchical relation between the three involved genome regions that are distinguished by expression mechanism. In the order of size increase these regions predominantly control genome replication, genome expression, and virus dissemination, respectively. In contrast to the observed directionality in the evolutionary dimension these fundamental biological processes cooperate bi-directionally on a functional level in the virus life cycle. Collectively, our findings suggest that genome architecture and the associated division of labor control genome size and may set its limits in RNA viruses.

Author Summary

RNA viruses include many major pathogens. Virus adaptation to their hosts is facilitated by fast mutation and constrained by small genome sizes, which are both due to the extremely high error rate of viral polymerases. Using an innovative computational approach we now provide evidence for additional forces that may control genome size and, consequently, affect virus adaptation to the host. We analyzed nidoviruses, a monophyletic group of viruses that populate the upper ~60% of the RNA virus genome size scale, evolved a conserved genomic architecture, and infect vertebrate and inverterbrate species. They include viruses with the largest known RNA genomes that exclusively encode a 3'-to-5'exoribonuclease, homolog of a canonical DNA proof-reading enzyme, which improves the replication fidelity. We show that the evolutionary space explored by these viruses exceeds that of the Tree of life for comparable protein datasets, although the time-scale of nidovirus evolution remains unknown. Extant nidoviruses with different genome sizes reached particular points on a common non-linear genome expansion trajectory. This trajectory may be shaped by the division of labor between open reading frames that predominantly control genome replication, genome expression, and virus dissemination, respectively. Ultimately, genomic architecture may determine the observed limit of genome size in contemporary RNA viruses.

Introduction

Genome size is a net result of evolution driven by the environment, mutation, and the genetics of the organism[308,442]. Particularly, mutation rate is a powerful evolutionary factor[116]. The relation between mutation rate and genome size is inversely proportional for a range of life forms from viroids to viruses to bacteria, and it is slightly positive for eukaryotes, suggestive a causative link[155,307,431]. The genome size of RNA viruses is restricted to a range of ~2-to-32 kb that corresponds to a very narrow band on the genome size scale from 1 kb to 10 Mb at which genome size increase is strongly correlated with mutation rate decrease[404]. This restricted genome size range of RNA viruses is believed to be a consequence of the lack of proof-reading factors resulting in a low fidelity of RNA replication[220,439]. In the above relation, mutation rate and proof-reading serve as a proxy for replication fidelity and genetic complexity, respectively. When combined, replication fidelity, genome size and genetic complexity form the unidirectional triangular relation that was postulated to lock these characteristics in low states in primitive self-replicating molecules[131]. The applicability of this trapping, known as the "Eigen paradox" [276], was also extended to RNA viruses[217]. Recent studies of the order *Nidovirales*, a large group of RNA viruses including those with the largest known genomes, provided strong support for the triangular relation and, unexpectedly, revealed a way of how the Eigen paradox could have been

solved by these viruses[336,432]. These advances established nidoviruses as a prime model for studying genome size evolution in RNA viruses.

The order *Nidovirales* unites viruses with enveloped virions and non-segmented single-stranded RNA genomes of positive polarity (ssRNA+), whose replication is mediated by cognate RNA-depended RNA polymerase (RdRp)[91,360]. The order includes four families - the *Arteriviridae* and *Coronaviridae* (including vertebrate, mostly mammal viruses), and the *Roniviridae* and provisional Mesoniviridae (invertebrate viruses). The unusually broad 12.7-to-31.7 kb genome size range of this monophyletic group of viruses includes the largest known RNA genomes that are employed by viruses from the families *Roniviridae* (~26 kb)[85] and *Coronaviridae* (from 26.3 to 31.7 kb)[90], collectively coined large-sized nidoviruses[174]. Viruses from the *Arteriviridae* (with 12.7-to-15.7 kb genome range)[140] and the recently identified Mesoniviridae (20.2 kb)[284] are considered small-sized and intermediate-sized nidoviruses, respectively. Nidoviruses share a conserved genomic architecture with multiple open reading frames (ORFs) that are flanked by two untranslated regions (UTRs)[49,84,98,336,500]. The two 5'-most ORFs 1a and 1b overlap by a few dozen nucleotides and are translated directly from the genomic RNA to produce polyproteins 1a (pp1a) and pp1ab, the latter involving a -1 ribosomal frameshift (RFS) event[55,366]. The pp1a and pp1ab are autoproteolytically processed to non-structural proteins (nsp), from nsp1 to nsp12 in arteriviruses and from nsp1 to nsp16 in coronaviruses (reviewed in [498]). They encode most components of the membrane-bound replication-transcription complex (RTC)[100,421,462] that mediates genome replication and the synthesis of subgenomic RNAs (known also as transcription)[409,450]. ORF1a encodes proteases for processing of pp1a and pp1ab (reviewed in [498]), trans-membrane domains/proteins (TM1, TM2 and TM3) anchoring the RTC[22,200] and numerous poorly characterized proteins. ORF1b encodes core enzymes of the RTC (see below). Other ORFs, whose number varies considerably among nidoviruses, are located immediately downstream of ORF1b and are expressed from 3'-coterminal subgenomic mRNAs (hereafter collectively referred to as 3'ORFs)[408]. They encode virion and, optionally, so-called "accessory proteins" (reviewed in [53,136,316]).

In addition to the genome architecture, nidoviruses share also an array (synteny) of 6 replicative protein domains. Three domains – an ORF1a-encoded protease with chymotrypsin-like fold (3C-like protease, 3CLpro) [13,27,179], an ORF1b-encoded RdRp[75,179,445] and a superfamily 1 helicase (HEL1)[178,212,417,419] that may form a part or entire protein released from pp1a/pp1ab – represent the most conserved enzymes (reviewed in [169]). For other proteins, a relationship may be established only for some lineages, mostly due to poor sequence similarity. Two tightly correlated properties separate large-sized and intermediate-sized nidoviruses from all other ssRNA+ viruses that form several dozens of families and hundreds species: the genome size exceeding 20 kb and the encoding of a RNA 3'-to-5'exoribonuclease (ExoN)[336]. The latter enzyme is distantly related to a DNA proofreading enzyme, and it is genetically segregated and expressed with RdRp and HEL1[323,432]. Based on these properties ExoN was implicated in improving the fidelity of RNA virus replication.

This hypothesis is strongly supported by an excessive accumulation of mutations in ExoN-defective mutants of two coronaviruses, mouse hepatitis virus[124] and severe acute respiratory syndrome coronavirus (SARS-CoV)[123] (for review see [99]), and the identification of the RNA 3'-end mismatch excision activity in the SARS-CoV nsp10/nsp14 complex[52]. In all likelihood, the on-going characterization of ExoN is expected to reveal the molecular mechanisms that control the fidelity of replication. Regardless of its details, the ExoN acquisition provides the most plausible explanation for the solving of the Eigen paradox with a single evolutionary event that likely liberated the ExoN-encoding nidoviruses for genome expansions beyond the limit observed by other non-segmented ssRNA+ viruses[174,336].

In this study we sought to gain insight into events that led to the emergence of the ExoN-encoding ancestor and for further expansion of the nidovirus genome to sizes threefold the average RNA virus genome size, hereafter referred to as the nidovirus genome expansion (NGE). We show that comparative sequence analysis of nidovirus families are complicated by huge evolutionary distances, at the scale of the Tree of life (ToL), that separate the most conserved proteins. To address this challenge, we exploited functional conservations in the genome architecture that could be established across the nidovirus genome in an alignment-free manner. Consequently we partitioned the genome into five spatially collinear regions. By employing a statistical framework we revealed non-linear, consecutive expansions of the three differentially expressed coding regions (ORF1a, ORF1b, 3'ORFs) that account for 95-99% of the genome. Importantly, these regions predominantly control, respectively, genome replication, genome expression, and virus dissemination, during the virus life cycle. The observed dynamics unveil an evolutionary pathway that accommodated both an enormous accumulation of mutations and virus adaptation to different host species. Our results also indicate that genome architecture and the associated division of labor control the expansion of RNA virus genomes and, contrary to the current paradigm exclusively focusing on replication fidelity, may determine the observed limit on RNA virus genome size.

Results

The scales of per-residue evolutionary change in nidoviruses and the Tree of life are comparable. Nidoviruses have evolved genomes in a size range that accounts for the upper ~60% of the entire RNA virus genome size scale and includes the largest RNA genomes[336]. How much did it take to produce this unprecedented innovation in the RNA virus world? This question could be addressed in two evolutionary dimensions: time and amount of substitutions. Due to both the lack of fossil records and the high viral mutation rate, the time scale of distant relations of RNA viruses remains technically difficult to study. Hence, we sought to estimate the amount of accumulated replacements in conserved nidovirus proteins

and to put it into a biological perspective by comparing it with that accumulated by proteins of cellular species in the ToL.

To this end, we used a rooted phylogeny for a set of 28 nidovirus representatives (Table S1), which is based on a multiple alignment of nidovirus-wide conserved protein regions in the 3CLpro, the RdRp and the HEL1, as described previously[336]. The 28 representatives cover the acknowledged species diversity of nidoviruses with completely sequenced genomes[85,90,140,284] and include two additional viruses. For the arterivirus species Porcine reproductive and respiratory syndrome virus we selected two viruses representing the European and North American types, respectively, because we observed an unusually high divergence of these lineages; for the ronivirus species Gill-associated virus we selected two viruses representing the genotypes gill-associated virus and yellow head virus, respectively, because these viruses showed a genetic distance comparable to that of some coronavirus species (CL & AEG, in preparation). The nidovirus-wide phylogenetic analysis consistently identified the five major lineages: subfamilies *Coronavirinae* and *Torovirinae*, and families *Arteriviridae*, *Roniviridae* and Mesoniviridae. The root was placed at the branch leading to arteriviruses (Fig. 1A) according to outgroup analyses[336]. Accordingly, arteriviruses with genome sizes of 12.7 to 15.7 kb are separated in the tree from other nidoviruses with larger genomes (20.2-31.7 kb).

We compared the evolutionary space explored by nidoviruses, measured in number of substitutions per site in conserved proteins, with that of a single-copy protein dataset representing the ToL[50] (Fig. 1B). Using a common normalized scale of [0,1], comparison of the viral and cellular trees and associated pairwise distance distributions revealed that the distances between cellular proteins (0.05-0.45 range) cover less than half the scale of those separating nidovirus proteins. (Fig. 1C). Unlike cellular species, nidoviruses form few compact clusters, which are very distantly related. The distances between nidovirus proteins are unevenly distributed: intragroup distances between nidoviruses forming major lineages are in the 0.0-0.25 range, while intergroup distances between nidoviruses that belong to different lineages are in the 0.55-1.0 range. The distances separating the intermediate-sized mesonivirus from other nidoviruses tend to be most equidistant, accounting for ~15% of all distances in the 0.55-0.85 range.

The scale of nidovirus genome size change is proportional to the amount of substitutions in the most conserved proteins. To explore the relation of genome size change and the accumulation of substitutions, we plotted pairwise evolutionary distances (PED) separating the most conserved replicative proteins (Y axis) versus genome size difference (X axis) for all pairs of nidoviruses in our dataset (Fig. 2). It should be noted that the observed genome size difference may serve only as a low estimate for the actual genome size change, since it does not account for (expansion or shrinkage) events that happened in parallel between two viruses since their divergence. The obtained 378 values

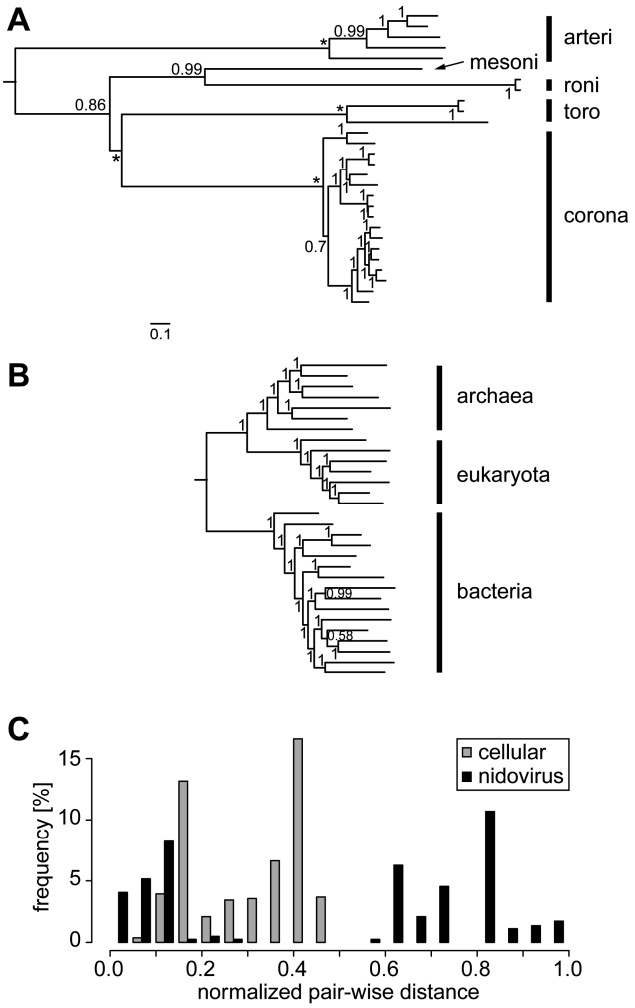


Figure 1. Phylogeny of nidoviruses in comparison to the Tree of life (ToL). Bayesian phylogenies of nidoviruses (A) and ToL (B) are drawn to a common scale of 0.1 amino acid substitutions per position. Major lineages are indicated by vertical bars and names; arteri: *Arteriviridae*, mesoni: Mesoniviridae, roni: *Roniviridae*, toro: *Torovirinae*, corona: *Coronavirinae*. Rooting was according to either (A) domain-specific outgroups[336] or (B) as described[50]. Posterior probability support values and fixed basal branch points (*) are indicated. The nidovirus and ToL alignments include, respectively, three enzymes and 56 single-gene protein families, 604 and 3336 columns, 2.95% and 2.8% gaps. For further details on the nidovirus tree see[336]. (C) Distributions of pair-wise distances for nidovirus and cellular single-copy conserved proteins according to the phylogenies in (A) and (B). The combined set of distances was normalized relative to the largest distance that was set to one.

are distributed highly unevenly, occupying the upper left triangle of the plot. Using phylogenetic considerations, four clusters could be recognized in the plot. Genetic variation within four major virus groups with more than one species (arteri-, corona-, roni-, and toroviruses) is confined to a compact cluster I in the left bottom corner (X range: 0.033-4.521 kb, Y range: 0.051-1.401). Values quantifying genetic divergence between major lineages are partitioned in three clusters taking in account genome sizes: large-sized vs. large-sized nidoviruses (cluster II, X: 0.002-5.433 kb, Y: 3.197-4.292), intermediate-sized vs. other lineages (cluster III, X: 4.475-11.494 kb, Y: 2.896-4.553), and small-sized vs. large-sized nidoviruses (cluster IV, X: 10.536-18.978 kb, Y: 4.159-5.088). Points in the clusters I, III and IV are indicative of a positive proportional relation between genome size change and the accumulation of replacements. The off-diagonal location of the cluster II can be reconciled with this interpretation under a (reasonable) assumption that the three lineages of large-sized nidoviruses expanded their genomes independently and considerably since diverging from the most recent common ancestor (MRCA). This positive relation is also most strongly supported by the lack of points in the bottom-right corner of the plot (large difference in genome size; small genetic divergence). Overall, this analysis indicates that a considerable change in genome size in nidoviruses could have been accomplished only over large evolutionary distances in the most conserved proteins.

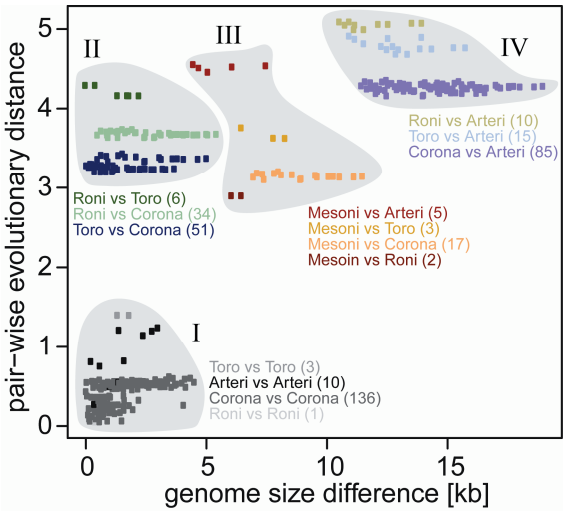


Figure 2. Relationship of evolutionary distance to genome size change in nidoviruses. Evolutionary distance (average number of substitutions per amino acid position in the conserved proteins) in relation to difference in genome size is shown for each pair (n=378) of the 28 nidovirus species. Points are colored according to pairs of major clades shown in Fig. 1A. The number of comparisons for each pair of clades is indicated by numbers in brackets. Points were grouped into clusters I (intra-lineage comparisons), II (large- vs. large-sized inter-lineage comparisons), III (intermediate-sized vs. others) and IV (small- vs. large-sized).

Only a fraction of genome size change may be linked to domain gain and loss. Next, we asked whether genome size change could be linked to domain gain and loss. We analyzed the phylogenetic distribution of protein domains that were found to be conserved in one or more of the five major nidovirus lineages[336]. Ancestral state parsimonious reconstruction was performed for the following proteins: ORF1b-encoded ExoN, N7-methyltransferase (NMT)[73], nidovirus-specific endoribonuclease (NendoU)[232,333], 2'-O-methyltransferase (OMT)[95,96], ronivirus-specific domain (RsD) (this study, see legend to Fig. S1), and ORF1a-encoded ADP-ribose-1"-phosphatase (ADRP)[129,375,402]. This analysis revealed that domain gain and loss have accompanied the NGE (Fig. S1 and Table S2). Particularly, genetically segregated ExoN, OMT and NMT (Fig. 3) were acquired in a yet-to-be determined order in the critical transition from small-sized to intermediate-sized nidovirus genomes. However, the combined size of these domains[336] accounts only for a fraction (49.7%) of the size difference (4,475 nt) between genomes of Nam Dinh virus (NDiV; 20,192 nt) and Simian hemorrhagic fever virus (SHFV), which has the largest known arterivirus genome (15,717 nt). The fraction that could be assigned to these and the three other protein domains is even smaller in other pairs of viruses representing different major nidovirus lineages (CL, AEG unpublished data). This analysis is also complicated by the uncertainty about the genome sizes of nidovirus ancestors that acquired or lost domains.

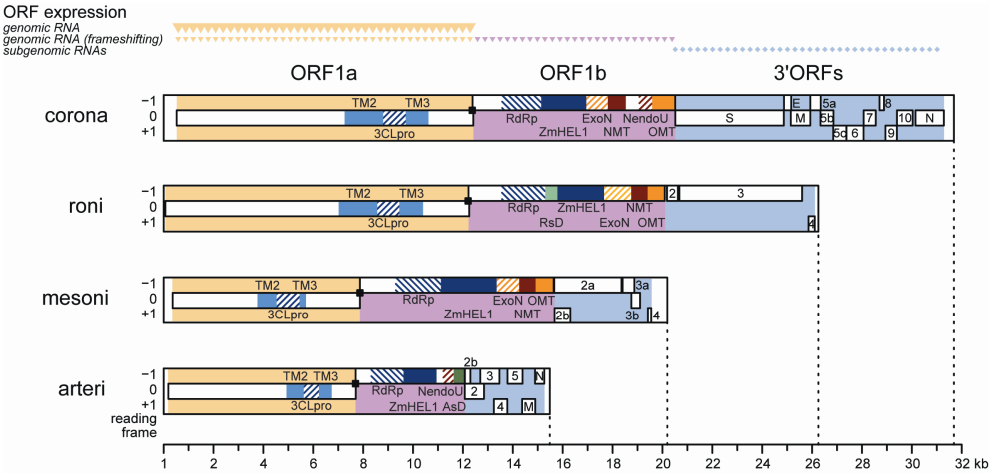


Figure 3. Genomic organization and expression, and key domains of four nidoviruses. The coding regions are partitioned into ORF1a (yellow), ORF1b (violet) and the 3'ORFs (blue), which also differ in expression mechanism as indicated on top. Black squares, ribosomal frameshifting sites. Within ORFs (white rectangles), colored patterns highlight domains identified in: all nidoviruses [TM2, TM3, 3CLpro, RdRp, and Zn-cluster binding domain fused with HEL1 (ZmHEL1)[461] - light and dark blue], large nidoviruses (ExoN, OMT - orange), certain clades (NMT, NendoU - red; ronivirus-specific domain (RsD) - light green; arterivirus-specific domain (AsD)- dark green). Genomic organizations are shown for Beluga whale coronavirus SW1 (corona), gill-associated virus (roni), Nam Dinh virus (mesoni), and porcine respiratory and reproductive syndrome virus North American type (arteri).

The nidovirus genome can be partitioned according to functional conservations in genome architecture. In order to gain further insight in the NGE dynamics, we had to analyze large genome areas in which homology signals are not recoverable in the currently available dataset because of both the extreme divergence of distant nidoviruses and a relatively poor virus sampling (Fig. 1). To address this challenge, we have developed an approach that establishes and exploits relationships between nidovirus genomes on grounds other than sequence homology. To this end, we partitioned the nidovirus genome according to functional conservations in the genome architecture, using results for few characterized nidoviruses and bioinformatics-based analysis for most other viruses (reviewed in [174]). With this approach, the genomes of all nidoviruses can be consistently partitioned in an alignment-free manner into five regions in the order from the 5'- to 3'-end: 5'-UTR, ORF1a, ORF1b, 3'ORFs, and 3'-UTR (Fig. 3). The 5'-UTR and 3'-UTR flank the ORFs area and account for <5% of the genome size in nidoviruses. The borders of the three ORF regions that overlap by few nucleotides in some or all nidoviruses were defined as follows: ORF1a: from ORF1a initiation codon to RFS signal, ORF1b: from RFS signal to ORF1b termination codon, and 3'ORFs: from ORF1b termination codon to the termination codon of the ORF that adjoins the 3'UTR.

It is noteworthy that the three ORF regions are of similar size but differ in expression mechanism (Fig. 3 top). Specifically, ORF1a is the first to be expressed by translation of the incoming virion RNA and, additionally, it encodes 3CLpro that mediates the release of mature proteins from the polyproteins pp1a and pp1ab. The expression of ORF1b, that follows, depends on the ORF1a region in three different ways: (i) the utilization of ribosomes that started translation on the ORF1a initiation codon; (ii) the use of the ORF1a/ORF1b RFS signal located upstream of the ORF1a termination codon; and (iii) the ORF1a-encoded 3CLpro. Finally, the expression of the 3'ORFs depends on products of the ORF1a and ORF1b to form the functional RTC for synthesizing subgenomic mRNAs that are translated to produce 3'ORF-encoded proteins[408]. Thus, ORF1a is the dominant region directly and indirectly controlling the expression of the entire genome.

The nidovirus genome expanded unevenly across three major coding regions. We then asked about how the different regions contributed to the genome expansion. We initially noted that the intermediate position of the mesonivirus between the two other nidovirus groups is observed only in genome but not region-specific size comparisons (Fig. 4). In the latter, the mesonivirus clusters with either small-sized (ORF1a and 3'ORFs) or large-sized (ORF1b) nidoviruses. This non-uniform position of the mesonivirus relative to other nidoviruses is indicative of a non-linear relationship between the size change of the complete genome and its various regions during the NGE. Accordingly, when fitting weighted linear regressions separately to the six datasets formed by nidoviruses with small and large genomes for three regions, support for a linear relationship was found only for the

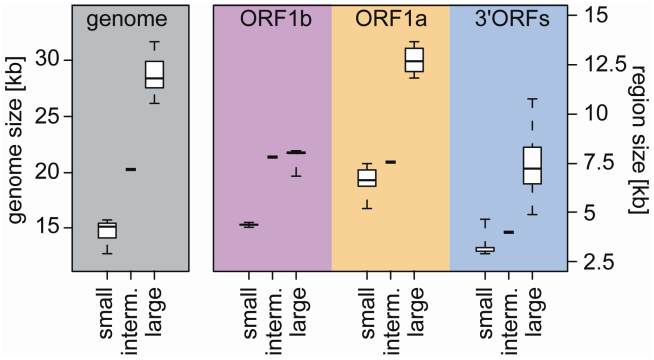


Figure 4. Nidovirus genome and region size differences. Shown are size distributions of genomes (left part) and the three genome coding parts ORF1a, ORF1b and 3'ORFs (right part) for five small-sized arterivirus species (small), 22 large-sized nidovirus species (large) and one intermediate-sized mesonivirus species (interm.). The distributions are represented by box-and-whisker graphs, where the box spans from the first to the third quartile and includes the median (bold line). The whiskers extend (dashed lines) to the extreme values.

3'ORF dataset of large nidoviruses; for all other regions a linear relationship was not statistically significant (Fig. S2). These results prompted us to evaluate linear as well as non-linear regression models applied to a dataset including all known nidovirus species (n=28) (Fig. 5). Two non-linear models were employed: third order monotone splines and a double-logistic regression. In the monotone splines, two parameters – the number and position of knots – determine the regression fit. We identified values for both parameters that result in the best fit (Fig. S3).

Using weighted r2 values, we observed that the splines model captures 92.9-96.1% of the data variation for the three ORF regions. This was a 5-22% gain in the fit compared to the linear model (75.9-90.8%) (Fig. 5). This gain was considered statistically significant (α=0.05) in two F-tests, a specially designed and standard one, as well as in the LV-test for every ORF region (p=0.018 or better) and, particularly, their combination (p=6.2e-5 or better) (Table 1). The splines model also significantly outperforms the double-logistic model (p=0.0011) (Table 1). These results established that the nidovirus genome expanded in a non-linear and region-specific fashion.

The three major coding regions expanded consecutively. Since each region expanded non-linearly during the NGE, so must the entire genome. Revealing its dynamic was our next goal. To this end, we analyzed the contribution of each of the five genomic regions to the overall genome size increase under the three models (Fig. 6 and Fig. S4). The top-ranked splines model (Table 1) predicts a cyclic pattern of overlapping wavelike increases of sizes for the three coding regions (the 5' and 3'UTR account only for a negligibly minor increase that is limited to small nidoviruses). Each of the three coding regions was found to

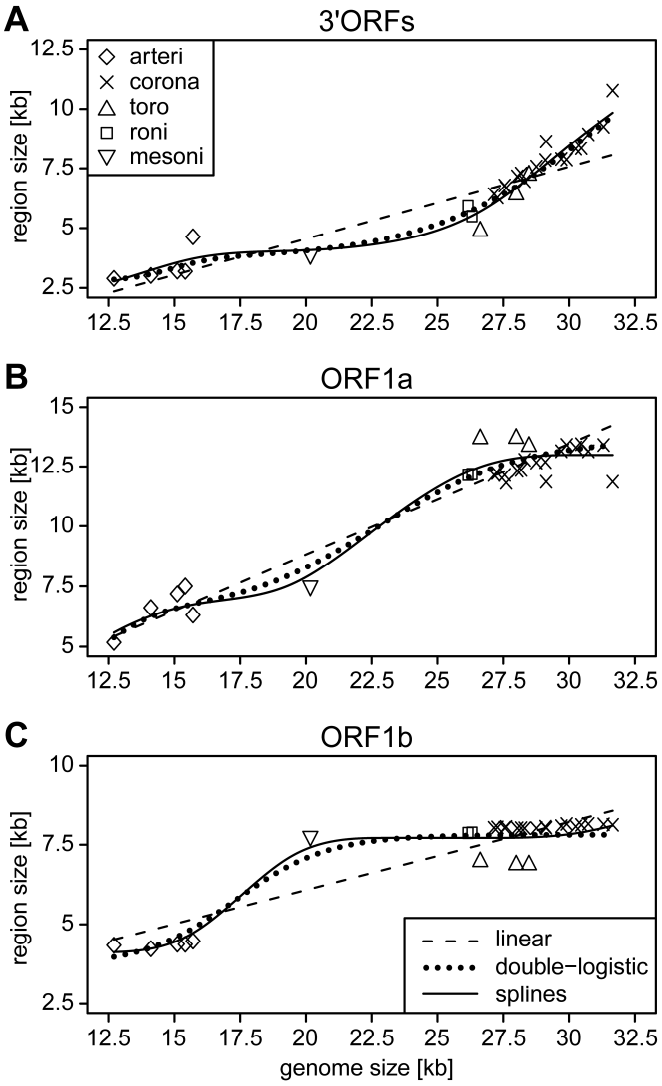


Figure 5. Relationship of sizes of three major coding regions and genome size in the nidovirus evolution. For 28 nidoviruses representing species diversity, absolute sizes of 3'ORFs (A), ORF1a (B), and ORF1b (C) are plotted against the size of the genome. Different symbols were used to group the viruses into five major phylogenetic lineages (see inlet in A). Results of weighted linear, double-logistic and 3rd order monotone splines[380] regression analyses are depicted. The three regression models (see inlet in C) fit the data with weighted r^2 values of 0.908 (linear), 0.948 (double-logistic) and 0.961 (splines) for ORF1a, 0.759, 0.900 and 0.929 for ORF1b, and 0.829, 0.950 and 0.955 for 3'ORFs. For fit comparison of regression models see Table 1.

Table 1. Comparison of regression models.

comparison[a]		test[b]	regression statistics[c]			
model A	model B		ORF1a	ORF1b	3'ORFs	total
linear	splines	F	0.0180*	0.0009*	0.0003*	5.2e-9*
linear	splines	F_{perm}	0.0008*	0.0028*	<1.0e-6*[d]	1.0e-6*
linear	splines	LV	0.0029*	0.0055*	0.0036*	6.2e-6*
linear	dlog	LV	0.0011*	0.0100*	0.0024*	6.5e-6*
dlog	splines	LV	0.0240*	0.0002*	0.20706	1.1e-3*

[a] linear regression model (linear); double-logistic regression model (dlog); 3rd order monotone splines regression model (splines)

[b] standard weighted F test (F); permutation F test (F_{perm}); a weighted version of a test to compare non-nested regression models (LV) as described in [286]

[c] shown is the probability that model A (null hypothesis) fits the data better than model B (alternative hypothesis); asterisks highlight significant values to reject the null in favor of the alternative hypothesis using a confidence level of 0.05; probabilities are calculated separately for ORF1a, ORF1b, 3'ORFs as well as the complete model combining the three coding plus the two UTR regions (total)

[d] non of the 1 million permutations resulted in an F larger than that of the non-permuted dataset

have been increased at different stages during the NGE (Fig. 6). A cycle involves expanding predominantly and consecutively the ORF1b, ORF1a and 3'ORFs region. One complete cycle flanked by two partial cycles are predicted to have occurred during the NGE from small-sized to large-sized nidoviruses. The complete cycle encompasses almost the entire genome size range of nidoviruses, starting from 12.7 kb and ending at 31.7 kb. The dominance of an ORF region in the increase of genome size was characterized by two parameters: a genome size range (X axis in Fig. 6) in which the contribution of a region accounts for a >50% share of the total increase, and by the maximal share it attains in the NGE (Y axis in Fig. 6). For three major regions these numbers are: ORF1b, dominance in the 15.8-19.3 kb range with 72.7% maximal contribution at genome size 17.6 kb; ORF1a, 19.6-25.9 kb and 83.0% at 22.4 kb; 3'ORFs, 26-31.7 kb and 89.8% at 29.4 kb (Fig. 6). Mesonivirus and roniviruses seem to have been "frozen" after the first (ORF1b) and second (ORF1a) wave, respectively. The third wave (3'ORFs) was due to the genome expansion of coronaviruses and, to a lesser extent, toroviruses (compare virus genome sizes on top with wave positions in Fig. 6).

Furthermore, the shapes of the three waves differ. The first one (ORF1b) is most symmetrical and it starts and ends at almost zero contribution to the genome change. This indicates that the ORF1b expansion is exceptionally constrained, which is in line with extremely narrow size ranges of ORF1b in arteri- and coronaviruses (with mean±s.d. of 4362±86 and 8071±50 nt, respectively; Fig. 4 and Fig. 6). The second wave (ORF1a) is tailed at the upper end and is connected to the ORF1a wave from the prior cycle. This ORF seems to have a relatively high baseline contribution (~20%) to the genome size change up to the range of coronaviruses. The third wave (3'ORFs) is most asymmetrical (incomplete),

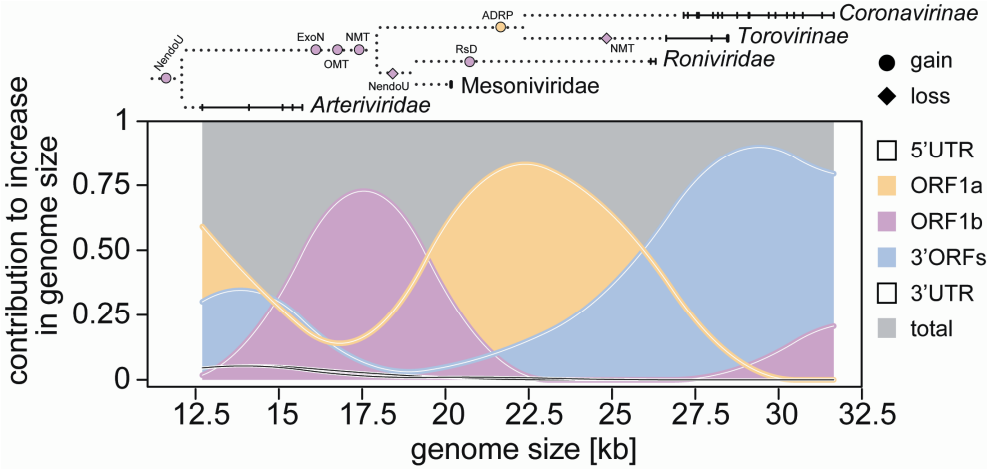


Figure 6. Region-specific, wavelike dynamics of the nidovirus genome expansions. Relative contributions of the genome regions ORF1a, ORF1b, 3'ORFs, 5'UTR and 3'UTR to the increase in genome size are calculated according to the splines regression and plotted on top of each other and against their sum=1. Solid horizontal lines and vertical bars on top: genome size ranges and samplings for nidovirus lineages indicated by names. Dotted lines: topology of major nidovirus branches. Selected domains gained (ExoN, OMT, NMT, RsD and ADRP, circles) and lost (NendoU and NMT, diamonds) are colored according to ORF in which they are encoded. See also Fig. 3, Fig. S1 and text.

as it only slightly decreases from its peak toward the largest nidovirus genome size at which this region remains the dominant contributor (~77%).

One partial cycle, preceding the complete one, is observed inside the genome size range of arteriviruses and involves the consecutive expansions of ORF1a and 3'ORFs, respectively. Also the main, but still very limited contributions of 5'- and 3'-UTRs (<6%) are observed here. The start of another incomplete cycle, involving the expansion of ORF1b and overlapping with the complete cycle, is observed within the upper end of coronavirus genome sizes.

Discussion

In this study we provide, for the first time, a quantitative insight into the large-scale evolutionary dynamics of genome expansion in RNA viruses. We analyzed nidoviruses, a monophyletic group of RNA viruses that populate the upper ~60% of the RNA virus genome size scale and include viruses with the largest known RNA genomes. Nidoviruses infect a broad range of different hosts including vertebrate and inverterbrate species and we now show that the evolutionary space explored by these viruses exceeds that of the ToL for comparable protein datasets. We exploited functional conservation in the genome

architecture in nidoviruses to partition their genomes in five spatially collinear regions. Using a complex statistical framework we reconstructed a non-linear trajectory of region-specific size increase that captured >92% of data variation. This trajectory may be shaped by the division of labor[442] between ORFs that predominantly control genome replication, genome expression, and virus dissemination, respectively. Combined, our results reveal that the genomic architecture severely constrains the NGE. Ultimately, it may determine the observed limit of genome size in contemporary RNA viruses.

Nidoviruses offer the best model for studying the control of RNA genome size. Genome size evolution in RNA viruses, unlike that of DNA-based life forms, has received relatively little attention from the research community. Several reasons may have contributed to this development. The narrow one-order range of small genome sizes that is compatible with the documented extremely high mutation rate[404] might have been perceived as evidence for the lack of meaningful genome size dynamics in RNA viruses. Even if there was any dynamics, its reconstruction could be considered challenging if not impossible to address, since evolutionary signals between distant lineages deteriorate profoundly due to the high mutation rate[220,489]. Consequently, the genome size increase in RNA viruses has so far been associated only with two trends to our knowledge: a concomitant increase of the average size of replicative proteins[33] and a reduction of genome compression measured by gene overlap[34].

In this respect, nidoviruses, which are often regarded an “exception” among RNA viruses[33,217], offer some unique opportunities for studying the evolution of RNA genome size. The genome size of nidoviruses is from ~20-to-200% larger than the “average” 10 kb RNA virus genome. Since nidoviruses form a monophyletic group and show a relatively large protein domain complexity, evolutionary analyses could be pursued.

Our results show that it took a considerable amount of evolutionary work in the most conserved proteins before a noticeable expansion of the nidovirus genome could be detected (Fig. 2). (In other, less conserved proteins the substitution rate is expected to be (much) larger). That relation is in line with an observation that nucleotide substitutions are on average four times more common than insertions/deletions in RNA viruses[404]. Whether this genome size increase also improves virus fitness and could determine the direction of evolution remains to be answered. In this respect we notice that viruses with larger genomes, compared to their small-sized cousins, could be expected to employ a more sophisticated repertoire of proteins for interacting with the host. It is also apparent that large-sized nidoviruses, unlike RNA viruses with smaller genomes, may afford both the acquisition and loss of an ORF as a matter of genome variation. Indeed, SARS-CoV adaptation to human and palm civets was accompanied with a large deletion in the ORF8-ORF10 area[193], and ORF gain/loss was documented in the recent evolution of other coronaviruses[68,302] (for review see [170]). Thus, large genomes could provide nidoviruses with an expanded toolkit to adapt upon crossing species barriers and to explore new niches in established hosts.

Inferring dynamics of genome size expansion in nidoviruses: viruses and protein domains. In our prior studies we already produced an unexpected insight into the control of genome size by identifying the ExoN domain in large-sized nidoviruses[432], a discovery that challenged a major paradigm of RNA virus biology - the universal lack of proof-reading during replication[220,439]. While this paradigm revision is getting support from experimental research[52,123,124,323], the recent discovery of a nidovirus in mosquitos, the mesonivirus NDiV, with a genome size in-between those of small-sized and large-sized nidoviruses, led to the proposal that 20 kb could be the genome size limit for non-segmented RNA viruses lacking ExoN (and proof-reading by implication)[336].

The identification of the 20 kb threshold poses questions about how nidoviruses have arrived at this threshold, crossed it, and expanded their genomes further. For addressing these questions we analyzed the entire ~19 kb genome size variation of nidoviruses (from 12.7 to 31.7 kb). We noted that only the lower ~20% and the upper ~30% of this range was sampled before the NDiV discovery. With the NDiV identification the ~50% non-sampled gap was split roughly in two halves, indicating that this sequence may provide a maximal information gain for analysis of the NGE (see also Fig. S2 in [336]). Indeed, an exceptionally large information value of the mesonivirus to this study is evident in many analyses (Figs. 3-6). On the other hand, the relatively strong impact of this single virus on the results may warrant an additional scrutiny to ensure the validity of conclusions. To this end, we list below other observations, in addition to the strong statistical significance (Table 1), that support the wavelike dynamics of the NGE. First of all, we note that a virus closely related to NDiV (called Cavally virus) was independently identified in a parallel study[500]. Both viruses share all properties that are critical for this study, including the size of genome and ORFs as well as the assignment of protein domains[284]. Second, these two mesoniviruses and the very distant roniviruses with large genomes form a monophyletic group (Fig. 1). This clustering correlates with common (molecular) properties, including the infection of invertebrate hosts and the lack of the NendoU domain, which distinguish mesoni- and roniviruses from other nidoviruses (Fig. S1) and could be expected to apply to other yet-to-be identified viruses of this group as well. Third, even if we restrict our analysis to small- and large-sized nidoviruses, differences between the size range of genomes and the three ORF regions are already apparent (Fig. 4). Particularly striking are the extremely constrained sizes of ORF1b in both arteriviruses and coronaviruses as well as an exceptionally large size range of 3'ORFs in large-sized nidoviruses. These constraints contribute prominently to the first and third wave, respectively, of the major cycle of the NGE (Fig. 6). Thus, the described dynamics of the region-specific genome size increase reflects properties of both mesoniviruses and other nidoviruses, and is expected to sustain upon future updates of virus sampling.

The available poor virus sampling limits the resolution of our reconstruction analysis of domain gain/loss during the NGE. For instance, the critically important acquisition of ExoN seems to be tightly correlated with those of two replicative

methyltransferases, NMT and OMT (Fig. S1). The fact that NMT and ExoN are adjacent domains in a single protein in coronaviruses (nsp14) and OMT resides nearby (nsp16) in pp1ab suggests a link between these domains and indicates that NMT and ExoN might have been acquired in a single event. Furthermore, NMT and OMT were shown to be essential for cap formation at the 5'-end of coronavirus mRNAs[73,95,96], with the OMT-mediated modification being important for the control of innate immunity[503]. These enzymes are yet to be characterized in other large-sized nidoviruses, and this characterization must reconcile the apparent lack of NMT in toroviruses[336] with its essential role in coronaviruses[73].

The ExoN acquisition is a hallmark of the first wave in the NGE because it is expected to have improved the replication fidelity and, thus, made further genome enlargements feasible. In contrast, no domain acquisition with a comparably strong biological rational could be identified for the second wave. Two aspects, both contrasting the first and second wave, are important to notice here. Firstly, while the first wave seems to reflect the genome expansion in a single ancestral lineage that might have given rise to all intermediate- and large-sized nidoviruses (founding event), the second wave is likely to encompass the expansions in several lineages that happened in parallel (Fig. S1b). Secondly, evolutionary relations of proteins in ORF1a (underlying the second wave) are not as extensively documented as those for ORF1b (underlying the first wave), since ORF1a proteins in nidoviruses have diverged far greater. Hence, the domain gain/loss description for the second wave is even less complete than that for the first wave. Most notable is the acquisition of ADRP (formerly X domain [180]) which seems to be part of the second wave in large-sized vertebrate nidoviruses (Fig. 6). This domain belongs to the macrodomain protein family with poorly understood function and a broad phyletic distribution in viruses and cellular organisms[357]. The ADRP was shown to have ADP-ribose-1''-phosphatase activity[375], bind poly-ADP-ribose[129], and its inactivation affected cytokine production in coronavirus-infected cells[137]. It was proposed to regulate RNA replication[432] and coronavirus pathogenesis[137], but its physiological function remains to be established. Unlike the first and second wave, the third one encompasses changes that predominantly happened during the radiation of a subfamily (*Coronavirinae*) rather than several families (Fig. 6); they are being analyzed in a separate study (CL & AEG, in preparation). Improved virus sampling in the future, especially in the genome size range around 20 kb, could be critical for the description of domain gain/loss in ORF1a and its refinement in ORF1b during the NGE (Fig. S1).

Genome architecture and division of labor may control dynamics of genome size expansion in nidoviruses. To analyze the dynamics of the NGE we exploited regional conservation of the expression mechanisms of ORFs in the nidovirus genome. This conservation has no parallel in the cellular world given the enormous accumulation of mutations it accommodated. It was established by combining results of comparative sequence analysis with those obtained by experimental characterization of few selected nidoviruses, mostly representing artriviruses and coronaviruses, the two polar groups in the

genome size dimension. Like with homology, functional considerations – in this case the roles of protein products in the viral life cycle and the order of ORF expression – were invoked to rationalize the observed conservation. Based on the available data, it could be argued that ORF1b, ORF1a, and 3'ORFs play predominant roles in genome replication, genome expression, and virus dissemination, respectively, in all nidoviruses. These three processes are essential for every virus and they form the backbone of the nidovirus life cycle (Fig. 7, bottom)[360]. ORF1b encodes the principal enzymes of RNA synthesis, e.g. RdRp, ORF1a controls the expression of all other ORFs by several mechanisms (see above), and the 3'ORFs encode the components of virus particles that are the principal vehicles of genome dissemination. The regional association of this dominant control of genome replication, genome expression, and virus dissemination may reflect the division of labor between the three non-overlapping coding regions of the genome in the nidovirus life cycle.

The cooperation between products of ORF1b, ORF1a, and 3'-ORFs is bidirectional in the nidovirus life cycle since the functioning of each region is critical for the two other regions. In contrast, the dynamics of genome expansion links these regions in the order ORF1b->ORF1a->3'ORFs (Fig. 7 top). It implies a predominantly unidirectional causative chain of regional expansion during the NGE that suggests a hierarchy of the three underlying biological processes. The association of the first wave of domain acquisitions with ORF1b attests for the universally critical role of replicative enzymes in the NGE beyond the 20 kb threshold that is observed by other ssRNA+ viruses (for discussion see [336]). Regardless in which order the OMT, NMT and ExoN loci were acquired, their products must have been adapted to the RTC whose enzymatic core is believed to be formed by ORF1b-encoded proteins[169,418,445]. Other, less conserved RTC components are encoded in ORF1a[96,200,229,371,409,491]. It is known that proteins encoded in ORF1a and ORF1b interact in coronaviruses[230,352,409] and some of these interactions, e.g. between nsp10 and nsp14 or nsp16, were shown to be essential for functioning of the ORF1b-encoded enzymes involved[51,52,74]. Accordingly, the RTC, already enlarged with the newly acquired ORF1b-encoded subunits, could have triggered and/or sustained expansion of ORF1a. Additionally, it may be prompted by the need to adapt the expression mechanisms for polyproteins 1a and 1ab, which were already increased in size and complexity in the ORF1b-encoded part. The final wave of expansion involving the 3'ORFs may be triggered by the need to adapt virus particles for accommodating the expanded genome[337]. During the NGE, a part of the newly acquired genetic material may have been adapted to facilitate both virus-host interactions[187,224,246,494] and inter-region coordination for the benefit of the processes they control and the life cycle[334]. For instance, in arteriviruses the ORF1a-encoded nsp1 is essential for subgenomic mRNA synthesis and virion biogenesis[332,454,455] and a role in transcription was proposed for an ORF1a-encoded domain of nsp3 in coronaviruses[210]. Thus, factors encoded by ORF1a and ORF1b might constrain the NGE by controlling the expression of the 3'ORFs region and/or the functioning of its products. This would explain why the 3'ORFs expansion could not have been possible before the expansion of ORF1a

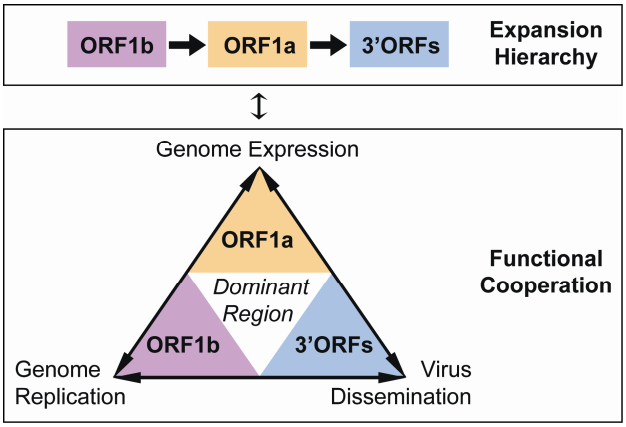


Figure 7. Hierarchy and cooperation in the nidovirus genome expansions. Functional and evolutionary relations between the three major coding regions of the nidovirus genome are depicted. For a brief description on the relationship between these three coding regions and the processes they dominate in the nidovirus life cycle, see text.

and ORF1b. By similar reasoning, an extremely tight control of the ORF1b size (Fig. 4) may set the ultimate size limit to the NGE. Finally, we note that the expansion order of the three coding regions matches their ranking according to sequence conservation, which is evident in the regional distribution of the nidovirus conserved domains (Figs. 2 and 3). This conservation is inversely proportional to the amount of accumulated substitutions, although quantitative characterization of the latter aspect is yet to be systematically documented. Genome changes due to regional-specific expansion and residue substitution may affect each other, and both may contribute to virus adaptation to the host.

Concluding Remarks and Implications. It is broadly acknowledged that extremely high mutation rates and large population sizes allow RNA viruses to explore an enormous evolutionary space and to adapt to their host[33,107]. Yet the low fidelity of replication also confines their evolution within a narrow genome size range that must affect their adaptation. Above, we presented evidence for a new source of constrains of genome expansion in RNA viruses by analyzing nidoviruses which include viruses with improved replication fidelity. In our analysis conserved genome architecture and the associated division of labor emerged as potentially powerful forces for selecting new genes and target genome regions during genome expansion. Importantly, the major diversification of nidoviruses by genome expansion must have started at some early point after the acquisition of ExoN[336]. From that point nidoviruses expanded their genomes in parallel in an increasing number of lineages, each of which may have acquired different domains in a same region. Extant nidoviruses of major lineages have very different genome sizes which we found to correspond to particular

points on the common region-specific genome expansion trajectory. The entire nidovirus (genome size) diversity may serve as a snapshot of different stages of the NGE. For viruses with largest genomes those with smaller genomes represent stages that they have passed in the NGE. For smaller genomes those with the larger ones represent stages that they have not reached in the NGE. It seems that the host may play a role in this process since ExoN-encoding nidoviruses that infect invertebrate are at the low side of genome size. For yet-to-be described nidoviruses, the genome expansion model can predict sizes of three coding regions by knowing only the genome size. The mechanistic basis of this fundamental relation can be probed by comparative structure-function analyses that should also advance the development of nidovirus-based vectors and rational measures of virus control. Thus, the wavelike dynamics model links virus discovery to basic research and its various applications.

This study indicates that genome size in RNA viruses may be restricted by the genome architecture in addition to the low fidelity of replication. Ultimately, these constraints may determine the upper limit of the RNA virus genome size. The reported data point to an important evolutionary asymmetry during genome expansion, which concerns the relation between proteins controlling genome replication, expression, and dissemination, and may be relevant beyond the viruses analyzed here.

Methods

Datasets. A dataset of nidoviruses representing species diversity from the three established and a newly proposed virus family was used (Table S1). A multiple alignment of nidovirus-wide conserved protein domains (28 species, 3 protein families, 604 aa alignment positions, 2.95% gap content) as described previously[336] formed the basis of all phylogenetic analyses. To put the scale of the nidovirus evolution into an independent perspective, we compared it with a cellular dataset previously used to reconstruct the Tree of Life, for which a concatenated alignment of single-copy proteins was used (30 species, 56 protein families, 3336 aa alignment positions, 2.8% gap content)[50]. The proteins used in the nidoviral and cellular datasets are the most conserved in their group and, as such, could be considered roughly equivalent and suitable for the purpose of this comparative analysis.

Phylogenetic analyses. Rooted phylogenetic reconstructions by Bayesian posterior probability trees utilizing BEAST[119] under the WAG amino acid substitution matrix[478] and relaxed molecular clock (lognormal distribution)[118] were performed as described previously[336]. Evolutionary pairwise distances were calculated from the tree branches. A maximum parsimony reconstruction of the ancestral nidovirus protein domain states at internal nodes of the nidovirus tree was conducted using PAML4 [487]. The quality of ancestral reconstructions was assessed by accuracy values provided by PAML4. To correct for non-

independence of the sequences[146] we assigned relative weights to the 28 nidovirus species by using position-based sequence weights[209] that were calculated on the alignment submitted for phylogeny reconstruction. The weights were normalized to sum up to one and were used in regression analyses (see below). The sequence weights varied ~7 fold from 0.017 to 0.116. NDiV, which represents mesoniviruses, showed the largest weight of 0.116 that was distantly followed by those of the bafinivirus White bream virus (WBV; 0.075) and roniviruses (0.06 each); coronaviruses, making up the best-sampled clade, were assigned the lowest weights (0.017 to 0.028 each).

Statistical analysis of genome size change in nidoviruses. The genome of each nidoviruses was consistently partitioned into five genomic regions according to external knowledge (see Results). To model the contribution of each genomic region to the total genome size change, we conducted weighted regression analyses (size of a genomic region on size of the genome) using three models – a linear and two non-linear ones. Position-based sequence weights were used and a confidence level of $\alpha=0.05$ was applied in all analyses. The combined contributions of all genomic regions to the genome size change must obviously sum up to 100%. To satisfy this common constraint, in each analysis, regression functions were fitted simultaneously to sizes of the genomic regions by minimizing the residual sum of squares, thereby constraining the sum of all slopes to be not larger than one. The linear model assumes a constant contribution of each genomic region during evolution which was modeled via linear regions.

In the first non-linear model we applied third order monotone splines with equidistant knots[380]. We chose splines because of their flexibility and generality (we don't rely on a specific regression function). The monotonicity constraint was enforced to avoid overfitting which was observed otherwise, and third order functions were chosen to obtain smooth, second-order derivatives. We explored the dependence of the performance of the splines model on variations in two critical parameters, the number of knots and the start position of the first knot. These two parameters define a knot configuration and determine a partitioning of the data into bins. In the first test we evaluated five different configurations generating from three to seven knots. Configurations using eight or more knots resulted in some bins being empty and were therefore not considered. For each number of knots the position of the first knot and the knot distance were determined as resulting in that configuration for which the data points are distributed most uniformly among the resulting bins. The exception was the 3-knot configuration, in which the position of the second knot was selected as the intermediate position in the observed genome size range (22.2kb). Only configurations with equidistant knots were considered. All probed splines models were evaluated by goodness-of-fit values (weighted version of the coefficient of determination r^2). In the second test we evaluated the model dependence on the position of the first knot by considering all positions that do not result in empty bins for the optimal number of knots determined using the approach described above.

As another non-linear model we used a 7-parameter double-logistic regression function that mimics the splines model and more readily allows for biological interpretations. Since double-logistic regressions did not converge for the 5'- and 3'-UTRs, linear functions were used for these two genome regions instead.

Linear (null hypothesis) and splines (alternative hypothesis) regression models were compared using standard weighted F-statistics and a specially designed permutation test (see below). To exclude overfitting as the cause of support of the more complex models, we utilized a more sophisticated framework (LV-Test) for the comparison of non-nested regression models (linear vs. double-logistic and splines vs. double-logistic) as detailed in [286]. The test was further modified to include weighted residuals according to virus sequence weights that account for sequence dependence.

Since our null hypothesis (linear model) is at the boundaries of the parameter space, we developed a permutation test to further compare the linear and splines models. To this end, genome region sizes were transformed to proportions (region size divided by genome size), randomly permuted relative to genome sizes, and transformed back to absolute values. These transformations are compatible with the constraints of the null hypothesis and the requirement that region sizes have to sum to genome sizes. Weights were not permuted. The linear and splines models were fit to the permuted datasets and F-statistics were calculated as for the original dataset. The p-value of the test is the fraction of F-statistics of permuted datasets that are larger than the F of the original dataset. It was calculated using 1,000,000 permutations that were randomly sampled out of ~10^{29} possible permutations.

Finally, we analyzed the contribution of each genome region to the total change in genome size under the three regression models. The contribution of each region according to a model was calculated as the ratio of change in region size to change in genome size (first derivative of the regression function) along the nidovirus genome size scale. These region-specific contributions were combined in a single plot for visualization purposes.

To conduct all statistical analyses and to visualize the results we used the R package[377].

Accession numbers. Accession numbers of virus genomes utilized in the study are shown in Table S1.

Acknowledgments

We thank Igor Sidorov for discussions and together with Alexander Kravchenko and Dmitry Samborskiy for Viralis management. This research has received funding from the Program of Japan Initiative for Global Research Network on Infectious Diseases (J-GRID), MEXT, Japan, the European Union Seventh Framework Programme (FP7/2007-2013) under the

program SILVER (grant agreement no. 260644), the Netherlands Bioinformatics Centre (BioRange SP3.2.2), the Collaborative Agreement in Bioinformatics between Leiden University Medical Center and Moscow State University (MoBiLe), and Leiden University Fund.

Supporting Information

Table S1. Nidovirus representatives.

virus	virus abbreviation[a]	(sub)family	accession[b]
Nam Dinh virus	NDiV_01-03	Mesoniviridae	DQ458789
Gill-associated virus	GAV_96	*Roniviridae*	AF227196
Yellow head virus	YHV_98	*Roniviridae*	EU487200
White bream virus	WBV-DF24_00	*Torovirinae*	NC_008516
Equine torovirus	EToV-Berne_72	*Torovirinae*	X52374
Bovine torovirus	BToV-Breda1_79	*Torovirinae*	NC_007447
Human coronavirus 229E	HCoV-229E_65	*Coronavirinae*	NC_002645
Human coronavirus NL63	HCoV-NL63_02	*Coronavirinae*	DQ445911
Miniopterus bat coronavirus 1	Mi-BatCoV-1A_05	*Coronavirinae*	NC_010437
Rhinolophus bat coronavirus HKU2	Rh-BatCoV-HKU2_06	*Coronavirinae*	NC_009988
Miniopterus bat coronavirus HKU8	Mi-BatCoV-HKU8_05	*Coronavirinae*	NC_010438
Scotophilus bat coronavirus 512	Sc-BatCoV-512_05	*Coronavirinae*	DQ648858
Porcine epidemic diarrhoea virus	PEDV-CV777_77	*Coronavirinae*	NC_003436
Feline coronavirus	FCoV_79	*Coronavirinae*	NC_007025
SARS coronavirus	SARS-HCoV_03	*Coronavirinae*	AY345988
Tylonycteris bat coronavirus HKU4	Ty-BatCoV-HKU4_04	*Coronavirinae*	EF065505
Pipistrellus bat coronavirus HKU5	Pi-BatCoV-HKU5_04	*Coronavirinae*	EF065509
Rousettus bat coronavirus HKU9	Ro-BatCoV-HKU9_05	*Coronavirinae*	EF065513
Human coronavirus HKU1	HCoV-HKU1_04	*Coronavirinae*	AY884001
Human coronavirus OC43	HCoV-OC43_67	*Coronavirinae*	AY585228
Mouse hepatitis virus	MHV-A59_59	*Coronavirinae*	AY700211
Infectious bronchitis virus	IBV-Beaud_35	*Coronavirinae*	NC_001451
Beluga whale coronavirus SW1	BWCoV-SW1_06	*Coronavirinae*	EU111742
Equine arteritis virus	EAV-CW_96	*Arteriviridae*	AY349167
Simian hemorrhagic fever virus	SHFV_64	*Arteriviridae*	NC_003092
Lactate dehydrogenase-elevating virus	LDV-P_71	*Arteriviridae*	U15146
Porcine respiratory and reproductive syndrome virus, North American type	PRRSV-NA_95	*Arteriviridae*	AF176348
Porcine respiratory and reproductive syndrome virus, European type	PRRSV-LV_91	*Arteriviridae*	M96262

[a] acronym of virus name joined ("_") with sampling year or period for this virus
[b] Genbank/Refseq accession number

Table S2. Nidovirus ancestral protein domain reconstruction.

ancestral node[a]	protein domain[b]											
	NendoU		ExoN		OMT		NMT		ADRP		RsD	
nido (root)	1	1.000	0	0.576	0	0.576	0	0.645	0	1.000	0	1.000
arteri	1	1.000	0	1.000	0	1.000	0	1.000	0	1.000	0	1.000
large nido+mesoni	1	1.000	1	1.000	1	1.000	1	0.836	0	1.000	0	1.000
mesoni+roni	0	1.000	1	1.000	1	1.000	1	1.000	0	1.000	0	1.000
roni	0	1.000	1	1.000	1	1.000	1	1.000	0	1.000	1	1.000
corona+toro	1	1.000	1	1.000	1	1.000	1	0.836	1	1.000	0	1.000
toro	1	1.000	1	1.000	1	1.000	0	1.000	1	1.000	0	1.000
corona	1	1.000	1	1.000	1	1.000	1	1.000	1	1.000	0	1.000

[a] abbreviations: nidoviruses (nido), large and intermediate size nidoviruses (large nido), roniviruses (roni), mesoniviruses (mesoni), toro-/bafiniviruses (toro), coronaviruses (corona), arteriviruses (arteri).

[b] shown are the reconstructed state (presence, 1, or absence, 0) and its accuracy by decimal numbers in the range of [0.500-1.000]) at the respective ancestral node for six domains in a maximum parsimony analysis using PAML.

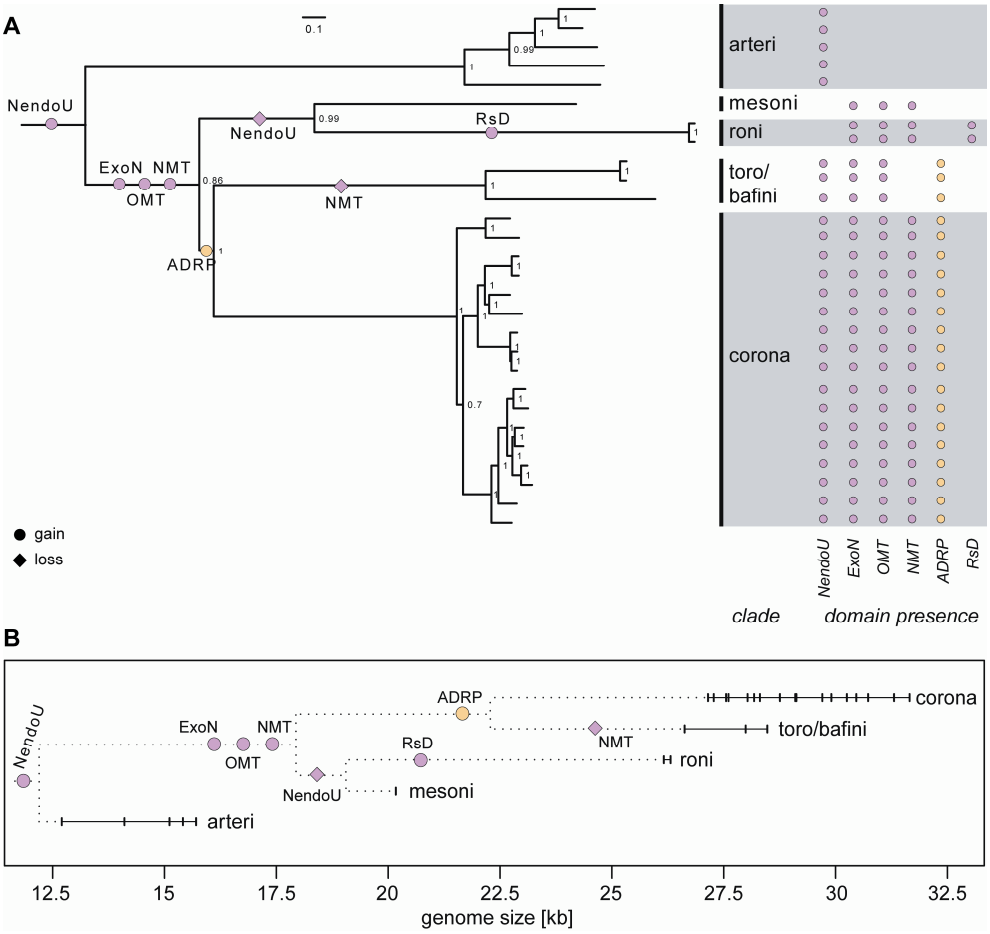


Figure S1. Gain and loss of selected ORF1a/ORF1b domains found in subsets of nidoviruses. (A) Distribution of six selected domains identified in ORF1a (one) and ORF1b (five) conserved in subsets of 28 nidovirus species (right part). One of the ORF1b-encoded domains (RsD) was identified in this study by inspection of the pp1b alignment as a ronivirus-specific insertion (163 aa) that is located between the conserved RdRp and ZmHEL1 domains (see Fig. 3). Colors indicate a domain's ORF location (purple for ORF1b, yellow for ORF1a). The left part shows predicted gain (circles colored according to its ORF location) and loss (colored diamonds) events at internal branches of the nidovirus phylogeny[336]. Nidovirus ancestral domain compositions were reconstructed utilizing a maximum parsimony analysis implemented in PAML4. Support values are shown in Table S2. (B) The nidovirus phylogeny was mapped on the genome size scale (dotted lines). Individual genome sizes of 28 nidovirus species are shown by vertical dashes and the size range within major lineages by horizontal solid lines. Internal nodes in the tree were arbitrarily placed at half the distance of adjacent branching events connecting two lineages while observing the original topology of the phylogeny. Predicted domain gain/loss events are highlighted as in (A).

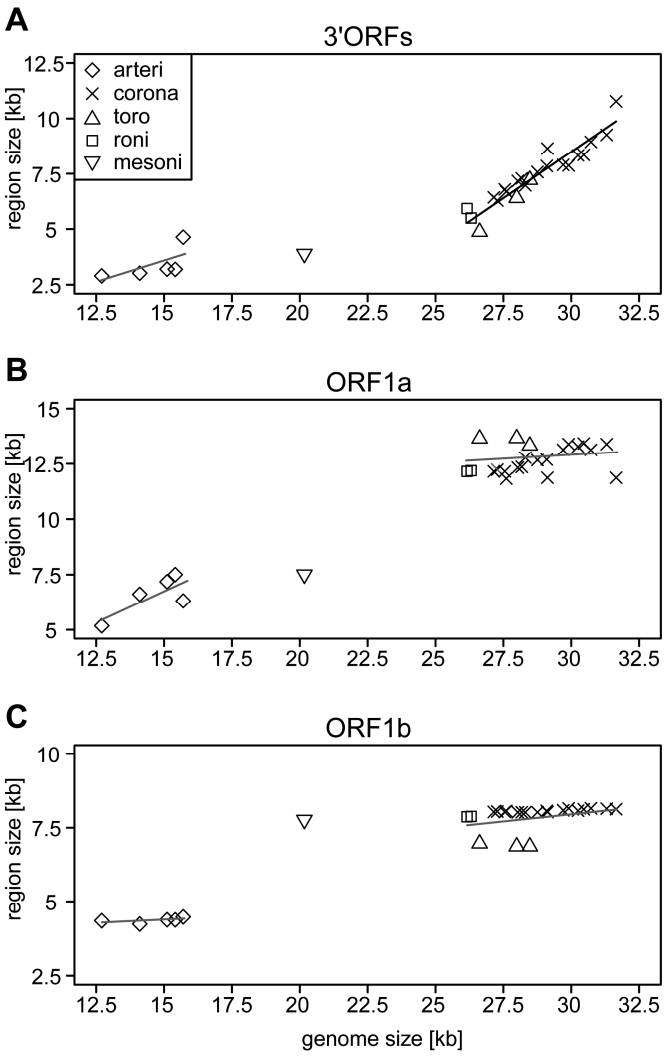


Figure S2. Clade-specific relationship of sizes of three major coding regions and genome size in the nidovirus evolution. For 28 nidoviruses representing species diversity, absolute sizes of 3'ORFs (A), ORF1a (B), and ORF1b (C) are plotted against the size of the genome. Different symbols were used to group the viruses into five major phylogenetic lineages (see inlet in A). Results of weighted linear regression analyses for small-sized (arteri) and large-sized nidoviruses (corona, toro/bafini, roni) are depicted. Regressions with a slope significantly different from zero are shown in black, non-significant ones in grey. The linear regressions fit the data with p=0.11, r^2=0.62 (arteri) and p=0.45, r^2=0.03 (corona, toro/bafini, roni) for ORF1a, p=0.33, r^2=0.31 and p=0.1, r^2=0.13 for ORF1b, and p=0.21, r^2=0.45 and p=6e-11, r^2=0.89 for 3'ORFs. The only significant correlation was observed for 3'ORFs of nidoviruses with large genomes (A) where the regression line showed a slope of 0.84 (±0.07 s.e.).

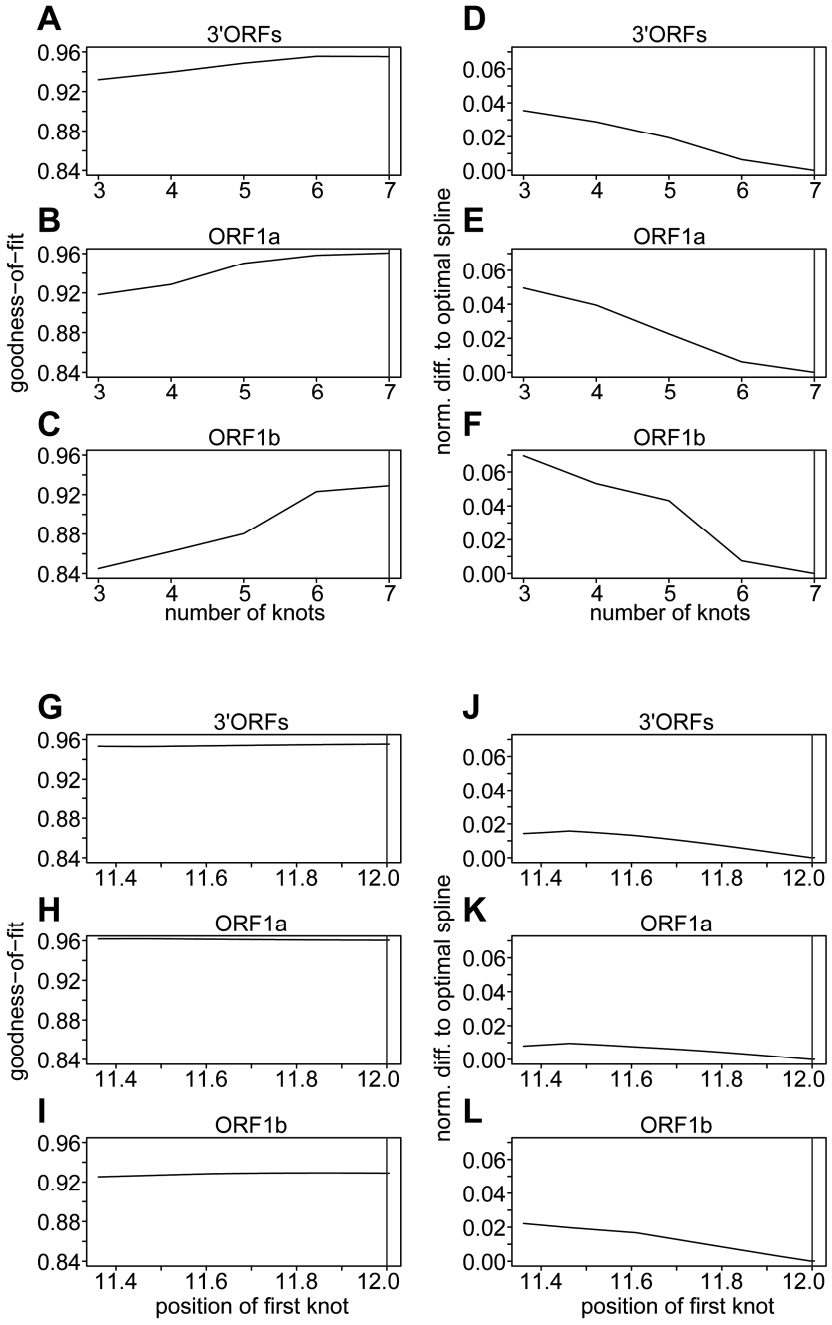

A
3'ORFs
B
ORF1a
C
ORF1b
goodness−of−fit
number of knots
D
3'ORFs
E
ORF1a
F
ORF1b
norm. diff. to optimal spline
number of knots
G
3'ORFs
H
ORF1a
I
ORF1b
goodness−of−fit
position of first knot
J
3'ORFs
K
ORF1a
L
ORF1b
norm. diff. to optimal spline
position of first knot

Figure S3. Sensitivity of the splines regression model to the number of knots and the position of the first knot. Shown are goodness-of-fit in form of weighted r^2 values (A-C, G-I) and sensitivity on the resulting regression curve (D-F, J-L) for different number of knots in the range of 3 to 7 (A-F) and different positions of the first knot (G-L) for the 3'ORFs, ORF1a and ORF1b genome regions. The best fit was obtained for the 7-knot configuration for all three regions (A-C). Hence, the 7-knot configuration was selected as the optimal one. We have also calculated a difference between other splines models compared to the optimal knot number by calculating the absolute difference of the regression curves of two configurations normalized to the size range of observed values (e.g. size ranges of ORF1a, ORF1b or 3'ORFs). This difference was in the range of 1-7% and increased with decreasing knot number in all three regions (D-F); it could be viewed as the loss of fit relative to the 7-knot configuration. Also, we calculated the model dependence on the position of the first knot by evaluating all positions that do not result in empty bins for the 7-knot configuration, which was found to be in the range from 11.4 to 12.0 kb (G-I). There was virtually no dependence of the position of the first knot and the goodness-of-fit (G-L); we selected the position that is closest to the minimal genome size. The knot number (k=7) and position of the first knot (at 12kb resulting in a knot distance of 3.7kb) used in the main calculation are indicated by green vertical lines.

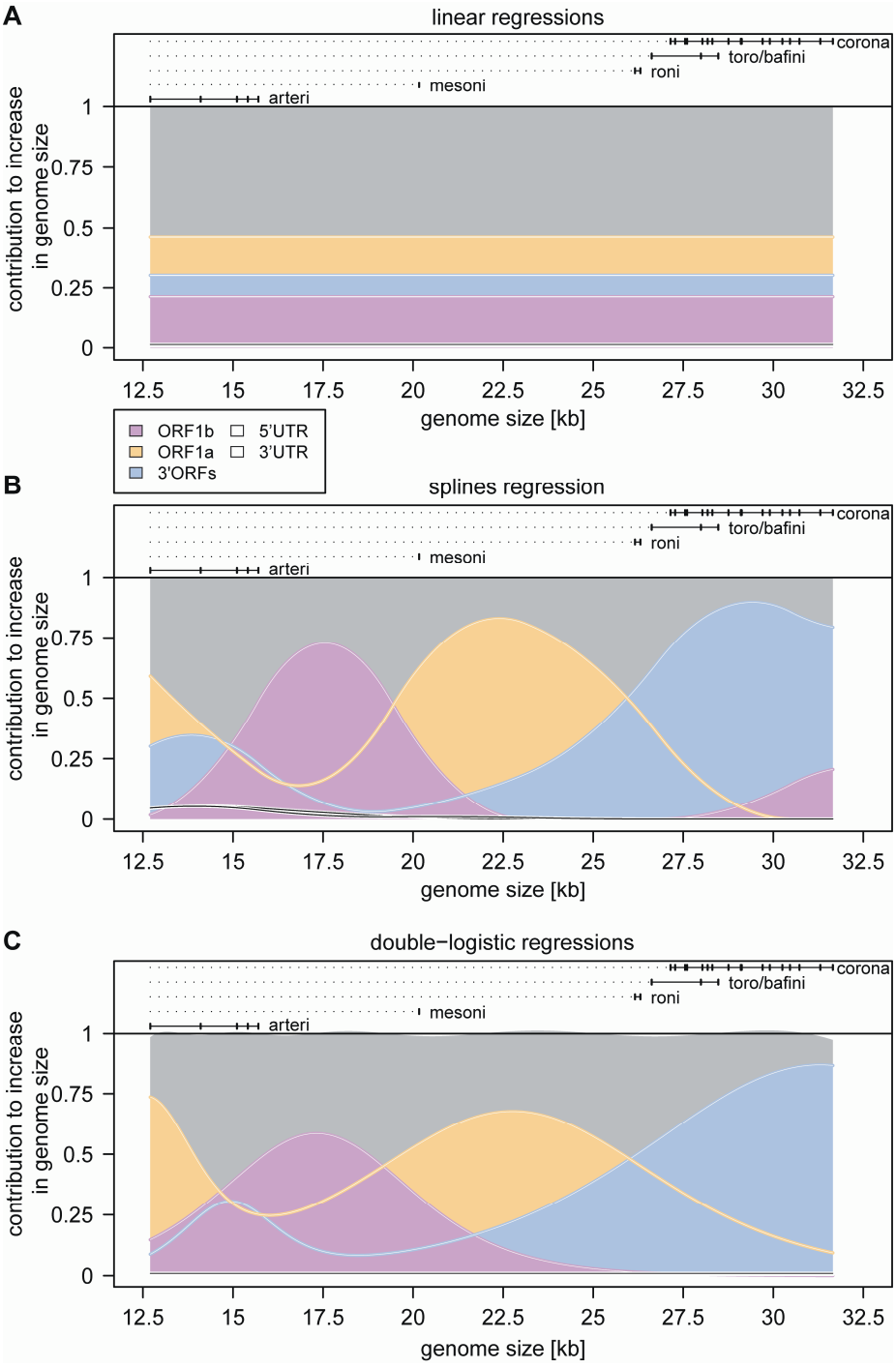

A
linear regressions
corona
toro/bafini
roni
mesoni
arteri
contribution to increase in genome size
1
0.75
0.5
0.25
0
12.5
15
17.5
20
22.5
25
27.5
30
32.5
genome size [kb]
ORF1b
ORF1a
3'ORFs
5'UTR
3'UTR
B
splines regression
C
double−logistic regressions

Figure S4. Modeling contribution of ORF1a, ORF1b, 3'ORFs, 5'UTR and 3'UTR to the nidovirus genome expansion. Relative contributions of ORF1a (yellow), ORF1b (purple), 3'ORFs (blue), and 5' and 3'UTR (black) to the increase in genome size are plotted on top of each other and against their sum=1 (grey) for the linear (A), the splines (B) and the double-logistc (C) regression model. Relative size contributions were calculated based on the regression curves fitted to the five genome parts for a dataset of 28 nidoviruses representing species diversity. Solid horizontal lines and vertical bars on top: genome size ranges and virus samplings for arteri-, corona-, toro-/bafini-, roni- and mesoniviruses. Under the linear model (which was statistically rejected in favor of the non-linear models), the contribution of each region to the genome size change is constant by definition. The ORF1a region accounts for most change (46.3%), followed by 3'ORFs (30.2%), ORF1b (21.3%), 5'UTR (1.3%) and 3'UTR (0.8%). In contrast, the splines and double-logistic models predict a cyclic pattern of overlapping wave-like increases of sizes for the three ORFs regions, with maximal contributions of 72.7%, 83.0% and 89.8% for ORF1b, ORF1a and 3'ORFs, respectively (see also main text). Highly similar cyclic and wave-like patterns of region expansions are predicted by the double-logistic model that mostly differs in the amplitude and range of waves compared to those of the splines model. These similarities suggest that the double-logistic model might be an approximation of the monotone splines model facilitating biologically meaningful interpretations.

CHAPTER 7

Origin and Evolution of the *Picornaviridae* Proteome

Alexander E. Gorbalenya
Chris Lauber

In: Ehrenfeld E., Domingo E., Roos R.P.
The Picornaviruses (2010), pp. 253–270
ASM Press, Washington, DC

Introduction

Picornavirus proteins are involved in all stages of the virus life cycle, from the virus entry into the cell and uncoating, to genome translation and replication, to encapsidation of the newly synthesized genomes and virion release from the cell[5,378] (see [346]). They interact with each other and virus RNAs as well as with cellular proteins, polynucleotides, and membrane components in performing their functions to produce virus and secure its spread outside the cell. Each protein adopts a variation of a common or unique fold and plays a particular role in a carefully orchestrated interplay for the virus to proliferate. Our understanding of the intricacies of this molecular machinery comes from studies involving functional and structural dissections of a few picornaviruses. Additional picornaviruses, whose number is growing, have been only poorly characterized, and the even much greater picornavirus diversity remains totally unexplored. For most of the known picornaviruses, genomes have been sequenced, enabling insight through comparative genomics. What is found to be conserved in sequences of all picornaviruses tends to be functionally and genetically essential in viruses that are studied in detail, implying a largely universal role for a conserved element in the picornavirus life cycle. Likewise, poorly conserved proteins tend to be dispensable in experiments and are involved in processes that modulate, often in a host-dependent manner, the picornavirus life cycle, which is driven by the key and most conserved proteins. This connection between functional, structural, and evolutionary dimensions forms a rational framework for model building in picornavirus research and serves for the dissemination of accumulated knowledge, both for established and for newly sequenced picornaviruses.

From a broad evolutionary perspective, picornaviruses form a phylogenetically compact family of viruses that infect vertebrates. Thousands of picornaviruses isolated so far can be grouped into 28 monophyletic lineages recognized as separate species and that are further grouped in 13 clusters taxonomically known as genera (Fig. 1)[258]. (The above numbers are not definitive, as most recently described picornaviruses may form additional species and prototype novel genera[249,250,386]; many more may come to light in the future[469] [see [264]].) Picornaviruses in a single species have limited sequence variability and always share a common protein set. Viruses that belong to different genera may typically be distinguished by the presence of one or more unique proteins (molecular markers); within a genus only a few species have such a distinction.

Picornaviruses employ a variant of a genetic plan common for a vast group of positive-stranded RNA viruses infecting also plants and invertebrates and known as the Picornavirales order[287]. The picornavirus genetic plan includes a single open reading frame (ORF) that occupies ~90% of the 6.7- to 8.8-kb genome[480]. It is flanked by the 5′ and 3′ untranslated regions (UTRs) that regulate translation and replication of the genome. The ORF includes 9 to 13 domains arranged in a conserved order and synthesized as a single polyprotein that is autoproteolytically processed to mature products. Three major virion (capsid) proteins, VP2 (1B), VP3 (1C), and VP1 (1D), are encoded in the N-terminal part of

the polyprotein. They are followed by proteins controlling replication and expression of the genome, 2B, 2C, 3A, 3B, 3C, and 3D, all listed in the order from the N to C terminus. Immediately upstream of 1B, some picornaviruses encode a small (minor) capsid protein VP4 (1A), a leader (L) protein and, in a single cardiovirus species (typified by Theiler's mouse encephalomyelitis virus [TMEV]), a second L protein in an alternative reading frame (L*). Likewise, downstream of 1D, a 2A protein is commonly encoded. L, L*, and 2A proteins are implicated in virus-host interactions (see [264,346,482] for a more detailed account of the picornavirus genome organization, encoded proteins, and infectious cycle).

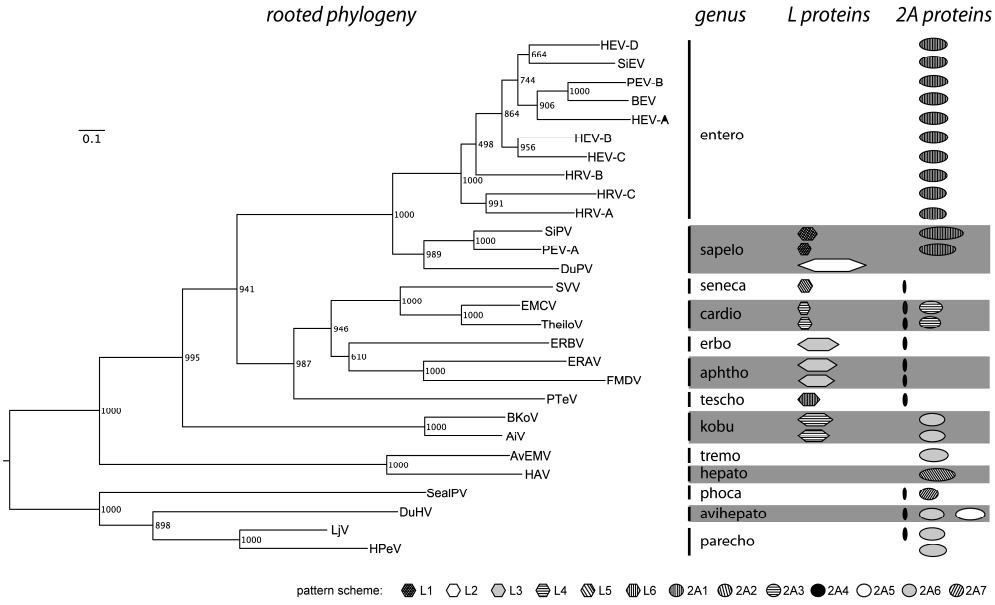


Figure 1. Phylogenetic tree of the *Picornaviridae* family. A phylogeny of 28 picornaviruses representing species diversity is shown. The maximum- likelihood tree is based on a multiple alignment of RdRps and was compiled using the PhyML program under the WAG amino acid substitution matrix and rate heterogeneity among sites (gamma distribution with four categories)[196,478]. A Bayesian reconstruction utilizing the BEAST software resulted in an identical topology. Numbers at branching points indicate bootstrap support values from 1,000 replicates. The scale of evolution in average number of amino acid substitutions per position is shown by the bar. The tree was rooted according to a separate phylogenetic analysis using nidovirus RdRps as an outgroup (data not shown). Picornavirus genera are indicated to the right of the phylogeny. For picornavirus species the presence of L and 2A proteins in polyproteins is depicted using rectangles of different shades. The widths of the rectangles are scaled proportionally to the size of L and 2A proteins. Homologous proteins are coded as described for Fig. 2, below. The viruses included are: HAV, avian encephalomyelitis virus (AvEMV), HPeV, LjV, DuHV AP, SealPV, porcine teschovirus (PTeV), FMDV SAT 2, ERAV, Theiler's-like virus of rats (TheiloV), encephalomyocarditis virus (EMCV), Seneca Valley virus (SVV), EERBV1, Aichi virus (AiV), bovine kobuvirus (BKoV), avian sapelovirus (DuPV), porcine sapelovirus (PEV-A), simian picornavirus 1 (SiPV), bovine enterovirus (BEV), simian enterovirus A (SiEV), HRV 30 (HRV-A), HRV-C, HRV-B, HEV-C, HEV-D, HEV 71 (HEV-A), HEV-B, and porcine enterovirus B (PEV-B).

In this chapter we briefly review our current understanding of the origin and evolutionary dynamic of the picornavirus proteome. There is a large body of literature on protein evolution recorded during picornavirus outbreaks and on picornavirus passaging in cells and animals in the absence or presence of a selective factor, e.g., a drug. These changes are limited in scale and restricted in time and can be defined as microevolution[107,108], which is outside the scope of this chapter (see [111,470] for further information). The evolution of picornavirus proteins is reviewed in reference [192]. Many concepts summarized and formulated in that review remain timely and are entertained below. The review can also serve as a reference point to observe advancements accomplished over the last 10 years. Discussions of evolution of all or some picornavirus proteins can be found in a number of other reviews[5,107,499].

Defining domain diversity in the picornavirus proteome

Before discussing picornavirus proteins, it is useful to recall that they were originally named without regard to evolutionary considerations, which is a common framework in contemporary studies. Accordingly, names were assigned to proteins based on either their electrophoretic mobility on gels (e.g., VP1 to VP4 for capsid proteins)[440] or by using a rational, genetically based nomenclature that divided the ORF into four consecutive nonoverlapping parts (L and 1 to 3), which were further split into 12 loci (L, 1A to 1D, 2A to 2C, and 3A to 3D)[398]. When comparative genomics of picornaviruses enabled evolutionary inferences, no large conflict was apparent between protein names and evolutionary relationships. The sole exceptions were the L and 2A proteins, which were found to belong to multiple protein families. To distinguish between evolutionarily unrelated L and 2A proteins in this chapter, each unique protein variant is assigned a number that follows either L or 2A. In total, six different Ls (from L1 to L6) and seven 2As (from 2A1 to 2A7) are currently recognized (Fig. 2) (unpublished data). They may be known under different names in other publications. The recognized varieties of L and 2A proteins are in no way definitive descriptions of the natural diversity of these proteins, whose full spectra may never be fully accounted as long as the picornavirus discovery effort is not extended to cover all vertebrate species.

The diversity of picornavirus proteins can be discussed and rationalized in the context of the picornavirus phylogeny (Fig. 1), the positions of respective genes on the genome (Fig. 2), and protein structures and their functions in virus reproduction. All these aspects are discussed in detail separately in many chapters in this book (mainly [115,261,264,291,313,346,397,463,482]). Here we will outline different properties of picornavirus proteins for the sake of defining their diversity before discussing the proteome evolution in some detail. A summary of the picornavirus proteome is given in Table 1.

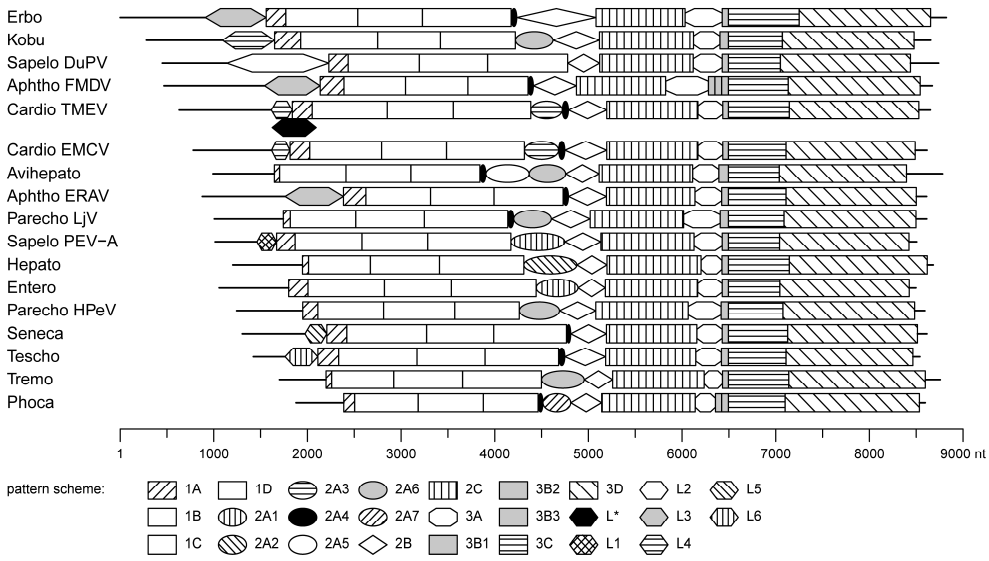


Figure 2. Polyprotein layout characteristics of genera or species of the *Picornaviridae* family. Genomic organizations for 17 picornaviruses of 13 genera are shown, describing all variants of the polyprotein domain architecture found for the *Picornaviridae* by 2009. The organizations were aligned at the 2B-2C border and are ranked in order of descending genome size. Mature proteins are depicted as different shaded rectangles (with the exception of 2A3 and 2A4 [NPGP] in cardioviruses, which are released as a fused product from the polyprotein), and UTRs are shown as solid horizontal lines. The identity of proteins can be determined using the legend at the bottom. Borders of proteins were identified using protein annotations for the most-well-characterized viruses, which were then applied to a family-wide polyprotein alignment, generated by using Muscle and curated manually with support of the Viralis software platform (Gorbalenya, unpublished). For the sake of this comparison, a region between a leader protein (where it is present) or the initiator codon (leaderless viruses) and 1B (VP2) was considered as 1A (VP4) in all viruses, although it is not produced in some viruses. For a discussion of the complexities of VP4 evolution, see the text. For TMEV two reading frames are shown (from top to bottom: 0 and +1 with respect to the start of the most upstream open reading frame), as it encodes an additional protein (L*) in the +1 frame.

Comparisons of polyproteins of picornaviruses have identified 1B, 1C, 1D, 2B, 2C, 3A, 3B, 3C, and 3D proteins in all picornaviruses sequenced to date. Conservation varies markedly between these functionally and structurally diverse proteins (see below). Other proteins, L, L*, 1A, and 2A, are not universally present in all picornaviruses. All proteins can be loosely placed in four groups based on functional and evolutionary considerations.

The first group comprises four capsid proteins, three of which, 1B, 1C, and 1D, have adopted different versions of an eight-stranded antiparallel β-barrel fold, dubbed a jelly roll[394]. Among RNA and DNA viruses of different families, this fold is most commonly used to build icosahedral capsids with different triangular numbers[399]. Picornaviruses use pseudo-T3 virion architectures[214,393]. This fold is also employed by cellular proteins[87,394]. The picornavirus capsid may include a small 1A (VP4) protein encoded upstream of 1B that is

involved in the interaction with the encapsidated RNA. The N-terminal myristoylation signal GXXXS/T is common in VP4 of many picornaviruses[78,310,312,355] whose apparent conservation is otherwise very limited[345] (Gorbalenya, unpublished). Also, in some picornaviruses, e.g., human parechovirus (HPeV)[437], Ljungan virus (LjV)[239], and duck hepatovirus type 1 (DuHV)[103,459], no 1A (VP4) protein has been identified.

The second group includes the 2C, 3C, and 3D proteins, all of which are key enzymes controlling replication and expression of the picornavirus genome; also, there is the 3B (VPg) protein, which is a substrate of 3D.

2C is a multifunctional protein involved in replication, membrane biogenesis, and virion uncoating. Bioinformatics-based analysis identified three domains in the 2C protein. Two α/α domains flank an ATP-binding/ATPase domain, adopting a variation of the α/β Rossmann fold, which is ubiquitous in the protein world[446,457]. The central ATPase domain belongs to the so-called helicase superfamily III (Hel SF3)[181]. It is characterized by three sequence motifs, A, B, and C, which are associated with enzymatic activities, ATPase, and presumably, helicase. Besides 2C of picornaviruses, the Hel SF3 includes (2C-like) proteins encoded by other viruses of the *Picornavirales* order, a few other single-stranded RNA (ssRNA) positive-strand viruses (ssRNA+), and small DNA viruses, as well as proteins of cellular origin. The Hel SF3 is related to a vast group of proteins known as the AAA+ ATPase superfamily, which adopt a ringshaped oligomer fold and are involved in myriad cellular processes[138]. The very N terminus of the N-terminal α/α domain of 2C was recognized as an amphipathic α-helix that may form a separate subdomain mediating interaction of 2C with membranes[354]. Its distant counterpart was identified in the N terminus of the unrelated NS5/NS5a proteins of flaviviruses, another ssRNA+ family[447].

3C is a cysteine proteinase, acting as the major enzyme mediating polyprotein proteolytic processing. It adopts the 12-stranded antiparallel two-β-barrel fold conserved in cellular serine proteases, with chymotrypsin as the prototype (reviewed in reference [430]). Many ssRNA+ viruses encode homologs (orthologs) of 3C protease, often known as 3C-like proteinases[185,401].

3D is the RNA-dependent RNA polymerase (RdRp) that mediates genome replication. It is related to a number of template-dependent polynucleotide polymerases, including RdRps of other RNA viruses, reverse transcriptases of viral and cellular origins, and DNA-dependent DNA polymerases[148,199,451]. These enzymes include a (palm) subdomain that adopts an RRM-like fold conserved among a number of functionally different proteins, including ribosomal proteins L7/L12 and S6, as well as the U1A splicing factor[199]. 3D is a primer-dependent RdRp that uridylates a very small 3B (VPg) protein to initiate RNA synthesis ([356]; reviewed in reference [149]). Functional counterparts of 3B have been identified in other ssRNA+ virus families in the *Picornavirales* order and also outside of the order[287,324]. They share no recognizable sequence similarity with 3B.

Table 1. The picornavirus proteome: function, structure, and evolution.

Protein		Function or involvement	Protein family	Range of the protein family		Origin[a] (source)	Reference(s)
				Viruses	Cells		
L	L1	Cysteine protease, polyprotein processing, anti-innate immunity	Papain-like	RNA	Yes	External	180,195
	L2	Zn binding, antiapoptotic, nucleocytoplasmic traffic, anti-innate immunity, translation shutoff	ORFan L2	Cardio-/Kobu-	No	De novo	369,388,392
	L3	Unknown	ORFan L3	PEV-A	No	De novo	274
	L4	Unknown	ORFan L4	DuPV	No	De novo	460
	L5	Unknown	ORFan L5	Seneca-	No	De novo	197
	L6	Unknown	ORFan L6	Tescho-	No	De novo	105
	L*	Antiapoptotic, proinflammatory	ORFan L*	TMEV	No	De novo	71,267
1A		Virion uncoating	?	Subset of picorna-	No	?	214,393
1B		Major capsid protein	Jelly roll	RNA/DNA	Yes	Ancestral	214,393
1C		Major capsid protein	Jelly roll	RNA/DNA	Yes	Ancestral	214,393
1D		Major capsid protein	Jelly roll	RNA/DNA	Yes	Ancestral	214,393
2A	2A1	Cysteine protease, polyprotein processing	Chymotrypsin-like	RNA	Yes	3C Duplicate	46,458
	2A2	Capsid-forming cofactor	ORFan 2A2	Hepato-	No	De novo	79,314,325
	2A3	RNA binding, translation, transcription shutoff	ORFan 2A3	Cardio-	No	De novo	12,171
	2A4	Separation of 2A and 2B proteins	NPGP	RNA	Yes	Ancestral?	113,348
	2A5	GTPase?	GTPase 1	Avihepato-	No	External	103,459
	2A6	Acyltransferase?	Permuted papain-like	RNA	Yes	External?	14,168,226
	2A7	GTPase?	GTPase 2	Phoca-	No	External	250
2B		Membrane anchoring	?	Picorna-	No	Ancestral?	92
2C		ATPase and predicted helicase activity, capsid assembly and uncoating, RNA synthesis, membrane remodeling	SF3 helicase	RNA/DNA	Yes	Ancestral	3,181,296,297,363,391, 446,457
3A		Membrane anchoring, inhibition of membrane/secretory traffic	?	Picorna-	No	Ancestral?	477
3B		Initiation of RNA synthesis	VPg	Picorna-	No	Ancestral?	356
3C		Cysteine protease, polyprotein processing, RNA binding	Chymotrypsin-like protease	RNA	Yes	Ancestral	10,15,47,186,347
3D		RNA-dependent RNA polymerase, replication	Palm-based polymerase	RNA/DNA	Yes	Ancestral	199,245,356

[a] Four types of domain sources for picornaviruses are considered: ancestral, de novo, duplication, and external: ancestral, descended from a domain present in the ancestral picornavirus; de novo, emerged by opening a portion of a reading frame for translation (overprinting) in a recent ancestor of the picornavirus; duplicate, acquired by a duplication of another picornavirus gene; external, acquired by horizontal transfer from a cellular or virus source outside of the picornavirus family.

A third group is formed by two proteins, 2B and 3A, that flank 2C in the polyprotein. Both mediate interactions with membranes to anchor the replicative complex (see [261,397,463]). They universally include a region enriched with hydrophobic amino acid residues that is the most characteristic sequence feature of these proteins[92,477], which are otherwise poorly conserved (see below and Fig. 3).

The L and 2A proteins form a fourth group. As was already mentioned, they are distinguished among picornavirus proteins for their unparalleled diversity of molecular forms. Seven 2A and six L proteins have been recognized in viruses of the established and provisional picornavirus species so far[392] (Fig. 1) (see [264,346]). It must be also noted that one virus, duck picornavirus (DuPV) of the *Sapelovirus* genus[460], may encode no 2A protein according to our analysis (Gorbalenya, unpublished). The L1 protein, which is a papain-like proteinase[195] encoded by three species of aphtho- and erboviruses, is the only L protein that has homologs encoded by other viruses and cellular organisms[180]. Other varieties of the L protein are encoded by genes found only in a single species, indicative of their recent origin. It is common to treat these de novo genes as belonging to ORFans, genes with no apparent homologs outside a restricted phylogenetic range due to their unique origin or fast evolution[424]. Interestingly, four molecular forms of L proteins have characteristic sequence signatures suggestive of Zn fingers ([70]; Gorbalenya, unpublished).

Some evolutionary characteristics of the 2A polypeptides parallel those of the L proteins, although important differences are also evident.

First of all, the 2A domain repertoire is dominated by proteins that have homologs outside picornaviruses rather than de novo proteins (Table 1). It includes only two ORFans, 2A3 and 2A2, in three species comprising two cardioviruses and hepatitis A virus (HAV), respectively. There are five molecular forms of 2A which have homologs outside the picornavirus family. 2A1, known as 2A cysteine chymotrypsin-like proteinase[28,46], is conserved in all 10 species of enteroviruses and two sapeloviruses. 2A6, originally designated as the Hbox-NC protein family[226], is a putative acyltransferase with a permuted papain-like fold[14,168] conserved in seven picornavirus species of one provisional and four established genera. 2A4, known otherwise - after the conserved sequence signature - as the NPGP or EXNPGP protein family[113,348], mediates polyprotein processing by a unique cotranslational mechanism[114]. It was identified in 10 species of eight genera. 2A5 and 2A7 are both putative GTPases, based on similarities to characterized cellular homologs; each was identified in a single picornavirus species[103,250,255,459].

Second, it was discovered that the 2A region can be multicistronic and accommodate up to three unrelated genes in a picornavirus. Recognized first in LjV[239], multidomain 2A organizations were later described in the newly discovered DuHV[103,255,459] and seal picornavirus type 1 (SealPV)[250]. In these three viruses each 2A protein may be released from the polyprotein as a separate protein moiety. In hindsight, a multicistronic organization of the 2A locus can also be recognized in two species of cardioviruses that are prototypes for the genus (Fig. 2). In these viruses, the 2A locus accommodates the 2A3 and

2A4 varieties that, unlike other 2A combinations, are expressed as a fused protein, known as cardiovirus 2A (G protein)[171]. Interestingly, 2A4 is part of all four known multicistronic 2A loci, although its relative position in the locus varies between LjV, DuHV, SealPV (5′-end proximal), and cardioviruses (3′-end proximal) (Fig. 2).

Recognizing protein conservation in the proteome

When discussing proteins of picornaviruses it is common, as has already been done in this chapter, to invoke protein conservation. There is a general consensus that 3D and 2C are the most conserved, 2A and L proteins are the least conserved, and all other proteins are distributed in between[350]. This notion is based on experience with comparative sequence and phylogenetic analyses of many researchers, as well as on some quantitative measurements, mostly restricted to closely related picornaviruses[63,225,351] or genera. Importantly, conservation may mean different things to different people. For instance, the mere fact that a protein is found in all picornaviruses could be considered sufficient to treat it as conservative, while the lack of a protein in one or more picornaviruses could lead to the opposite conclusion. Figures 1 and 2 compare the distributions of all protein varieties among picornavirus species in a systematic way to reveal ubiquitous (conserved) and lineagespecific (nonconserved) proteins.

Another way to look at protein conservation is through sequence motifs (or characteristic signatures or patterns) whose size, uniqueness, and number can be linked to the underlying sequence conservation. Sequence motifs represent a simplified description of regional position-specific variation in sequence alignments produced to maximize similarity of proteins[35,227]. Clusters of alignment positions with no or highly restricted variations are selected to define motifs; in proteins the respective amino acid residues may form active sites of enzymes and/or encompass structurally important elements. The link between motifs and sequence variations can be exploited in a systematic analysis of conservation in picornaviruses. Position-specific amino acid residue variations along the polyprotein can be plotted by utilizing a polyprotein-wide sequence alignment produced for different subsets of picornaviruses[59,345,350]. We have plotted position-specific similarities in polyproteins of the entire *Picornaviridae* family (Fig. 3). In this plot, regions of high similarity form peaks, which are separated by valleys corresponding to relatively low similarities. This profile is evidently informative only for loci encoding proteins found in all picornaviruses (ubiquitous proteins). For the L and 2A regions, accommodating multiple unrelated molecular forms, the similarity profile is mostly not informative (flat line) due to gap dominance in the majority of sequences in these regions of the polyprotein alignment. To reveal sequence conservation in these regions, similarity profiles can be plotted for separate 2A and L protein families represented by two or more sequences (Fig. 3, inserts). These separately built profiles are useful for assessing variations in a protein family but cannot be used for cross-comparisons between

different varieties of 2A and L proteins that are encoded by nonidentical subsets of picornavirus species.

Peaks in the polyprotein similarity plot correspond to motifs[60,184]. They can be used for verifying and expanding the motif assignments made in studies with limited sequence diversity and sampling. For instance, in 3Dpol, the A, B, C, E, F, and G motifs[184] that form the active site and reside predominantly in the palm subdomain occupy most of the major peaks. They match, in some cases with a deviation, the motifs assigned or used in earlier studies[245,268,367]. In 3Cpro, the four peaks correspond to sequence elements, including three catalytic residues, H, D/E, and C, and a major substrate-binding site (SB), containing the GXH signature[28,173]. Likewise, the A, B, and C motifs of the Hel SF3[181] comprise most of the large multipeak area of the 2C protein. Distinct peaks are also evident in several other proteins, most prominently in three capsid proteins, 1B, 1C, and 1D[345], as well as 3B (a Y peak comprising the G/AXYXG signature centered around the uridylated Tyr residue)[175,411].

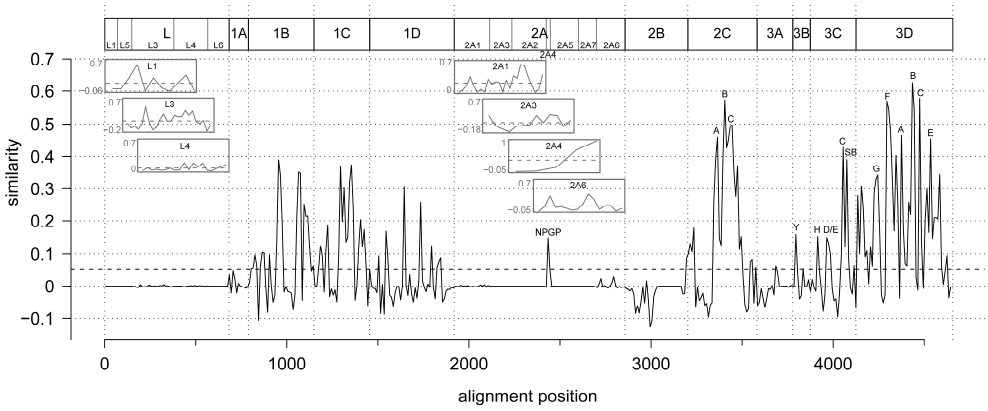


Figure 3. Polyprotein conservation of the *Picornaviridae* family. A plot of the conservation along the polyprotein alignment of 13 picornaviruses representing genus diversity is shown. The normalized similarity measure was compiled using the Bio3d package in R under the Blosum62 substitution matrix and a sliding window size of 10 amino acid positions[188,377]. The mean similarity of the polyprotein is indicated by the dashed horizontal line. On top, the positions of single protein alignments are highlighted by black rectangles and names with the same nomenclature as used for Fig. 2. For L and 2A proteins the positions of alignments for the different protein families (see also Table 1) are shown by grey vertical lines. The grey inserts represent separate conservation plots for the different L and 2A proteins that are expressed by at least two virus species. The following conserved sequence motifs are indicated at peaks of the similarity measure: NPGP cleavage motif in 2A4; 2C helicase motifs A, B, and C; 3B conserved Tyr (Y) nucleotidylated during priming in RNA synthesis; 3C protease catalytic His (H) and Cys (C), noncatalytic Asp/Glu (D/E) residues, and a substrate- binding motif (SB); 3D polymerase motifs A, B, C, E, F, and G.

Similarity peaks can be used to rank proteins according to their conservation. For instance, using the height of the tallest peak as a criterion (Fig. 3), the six most conserved proteins would be ranked in the order 3D, 2C, 3C, 1B, 1C, and 1D. Since proteins differ in size as well as numbers and shapes of similarity peaks, another measure of amino acid similarity that takes all these variables into account could be most inclusive. We designed such a measure and have called it a normalized similarity (Lauber and Gorbalenya, unpublished). It is calculated by adding up the similarities over all alignment positions of a genetic region or protein and dividing the obtained value by the number of alignment positions in the region or protein. This measure can be viewed as an integral indicator of protein conservation for a region or protein. According to this measure, five proteins, in descending order, 3D, 2C, 1C, 1B, and 3C, are the most conserved and form a separate group in the plot (Fig. 4). This ranking is close to that drawn from using the heights of the tallest peaks (see above). It is correlated with the key role of the five domains in the control of genome replication, expression, and encapsidation. Remarkably, these five proteins, compared to other proteins, are also distinguished by having the most restricted size variation (Fig. 4). These observations show that the evolution of key proteins of picornaviruses is most constrained in two dimensions that defined variation of amino acid sites and protein size, respectively. Mechanistically, constraints must have been imposed on accepting nonsynonymous replacements and in-frame insertion or deletions in these proteins.

Sources and mechanisms of innovation in the evolution of picornavirus proteins

The discussion above shows that replacement, insertion, and deletion of amino acid residues may fully account for the entire variation of the five most conserved picornavirus proteins. Excluding the N-terminal helix of 2C, discussed separately below, this notion is also supported by the lack of mosaic relationships between these and other proteins. Two processes, mutation and homologous recombination, have been shown to be involved in generating these changes in the most conserved proteins. Mutation is produced as a result of nucleotide misincorporation mediated by the RdRp during replication that may be translated in local changes in the protein (see [111,470]). In contrast, homologous recombination, in which virus progeny are generated by two parents that exchange homologous parts (see [6,426]), may affect a relatively large genome region encoding one or more proteins. Homologous recombination may require extensive base pairing to occur, which would restrict it to closely related viruses. Accordingly, homologous recombination has been implicated in generating intraspecies protein diversity, while mutation is a mechanism operating with no apparent phylogenetic or taxonomic barriers (see also [6,111,426,470]).

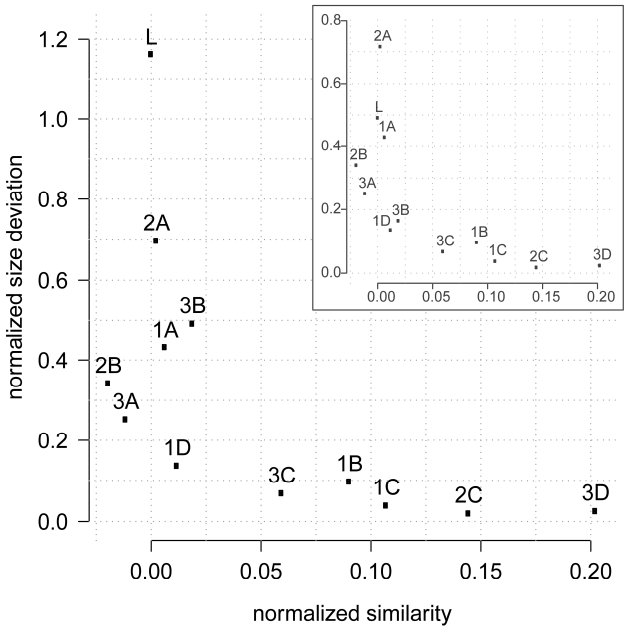


Figure 4. Protein conservation in the *Picornaviridae* family. For each of the different picornavirus proteins the mean normalized similarity (see Fig. 3 and text) is plotted against the length deviation, where the latter was compiled as the standard deviation divided by the mean length. For the main figure (in black), lengths of protein regions L, 1A, 1B, 1C, 1D, 2A, 2B, 2C, 3A, 3B, 3C, and 3D were used (allowing lengths of 0 in cases of absent proteins), whereas lengths in the inset plot (grey) are based on mature proteins (absent proteins were not counted). The same data set used for Fig. 3 was used here.

These processes contribute to the evolution of all proteins, regardless of the level of protein conservation[4]. They can also facilitate the proteome expansion by accepting nucleotide substitutions that open an additional region for reading by ribosomes. All ORFans identified in the L region of different picornaviruses could have emerged through converting a part of the 5′ UTR adjacent to the polyprotein ORF into the coding locus by mutation. Thus this conversion may not necessarily be accompanied by the loss of the original function, implying that the converted region could combine two roles: the original one as part of the 5′ UTR and a newly acquired role as a polyprotein domain. This region in picornaviruses is known to carry few essential RNA signals, as it typically separates the highly structured internal ribosome entry site from the initiator codon located downstream[8] (see [313]). Evidence for functional overlap of the 5′ UTR with the polyprotein ORF has also been reported for other viruses[495]. The reverse scenario—loss of an L protein due to mutation and/or nonhomologous recombination—seems to be equally plausible. The genetically engineered deletion of the L gene was largely tolerated by foot-and-mouth disease virus (FMDV)[365] and TMEV[266]. Also, an FMDV-based chimera carrying a noncognate L protein from Theiler

murine encephalomyelitis virus (TMEV) was found to be viable[364]. The loss of the L gene could have accompanied the emergence of the ancestor of leaderless enteroviruses, whose phylogenetic neighborhood includes exclusively Lcontaining lineages comprising seven genera (Fig. 1).

Another way of expanding the proteome without accompanying genome expansion is to open an alternative reading frame to the polyprotein ORF. Due to its redundancy, the genetic code can be used to maintain two or more ORFs in a genomic region (overlapping ORFs) at the expense of restricting varieties of codons used in the overlapping ORFs. The mutation process leading to the emergence of such alternative ORFs is called overprinting[252]. It is relatively common in RNA viruses, which evolve under tight constraints imposed on the genome size expansion by low-fidelity replication[34]. Overprinting is believed to generate de novo genes that, like many L genes discussed above, belong to ORFans[424]. The emergence of the L* gene, overlapping with the L gene and found exclusively in a subset of viruses of the TMEV species[71,267], was likely via overprinting. In line with the above hypothesis, the naturally leaderless poliovirus was artificially converted with few replacements in the 5′ UTR into a viable mutant carrying an L* gene overlapping with the polyprotein ORF[361]. On the other hand, it has been speculated that some varieties of 2A proteins found in many viruses could have evolved from ancestors encoded in alternative ORFs that may be of ancient origin[192].

The origin and evolution of the least-conserved proteins may also involve nonhomologous recombination (see [6,426]), which is a process of generating recombinant virus progeny by two parents that share no extensive homology. One parent, which could be called “minor,” contributes a nonhomologous gene to incorporate it as an insertion in the progeny genome most similar to the other, major, parent. A minor parent may be of cellular or viral origin, and it serves as an external unrelated source of genetic variability for progeny of the major partner. Special cases of nonhomologous recombination are gene duplication and loss in progeny of a single parent. In the case of gene duplication, a genetic locus is repeatedly copied, while gene loss is a result of skipping a genetic locus from copying; both are considered to be aberrations of template-mediated replication in picornaviruses. From the evolutionary point of view, gene duplication and loss can be considered an expansion of genetic variation by using the cognate template in an aberrant way.

As evident from the extreme sequence diversity of 2B and 3A proteins (Fig. 3), their evolution seems considerably less constrained than that of the most conserved proteins. Relatively high rates of mutation fixation and homologous recombination may fully account for generating this diversity, which does not include clearly recognizable unrelated molecular forms, like those found in L and 2A proteins.

Similar reasoning can be applied to explain the diversity of 1A (VP4) proteins. In this framework, picornaviruses with no identified VP4 could have evolved highly divergent variants of this protein that are covalently fused to the downstream 1B protein. Because they possess deviant properties, such variants may have eluded identification by bioinformatics

analyses and in structural studies. Indeed, picornaviruses, which are known to lack 1A, have a 1B that is N-terminally extended (Gorbalenya, unpublished). Alternatively, the above correlation could be fortuitous, and 1A might have been lost in the evolution of some picornaviruses.

Next to the most conserved proteins at the amino acid conservation scale is the capsid 1D protein (Fig. 4). Compared to its more conserved 1B and 1C paralogs, 1D is most exposed in virions, which allows it to interact with antibodies and receptors; this exposure could be a major reason for its evolution being less constrained[26]. Several picornaviruses, e.g., FMDV and strains of human enterovirus B species, have evolved, likely independently, a receptor-binding RGD signal in different places of the C-terminal region of 1D[26,48,208,315,468,479]. Given the small size of this signal and its nonessential role in the replicative cycle, it could have originated by mutation. Alternatively, a nonhomologous recombination with an unknown cellular or viral protein employing the RGD tripeptide might have been involved. These scenarios are not mutually exclusive, and both might have contributed to the diversity of the RGD-containing proteins in picornaviruses[166].

Nonhomologous recombination can be invoked as the most likely mechanism responsible for the origin of picornavirus proteins with limited phyletic distribution among picornaviruses that have either paralogs in picornaviruses or homologs of other, virus or cellular, origins. This list includes L1, 2A1, 2A5, 2A7, and 3B proteins (Table 1). The L1, 2A5, and 2A7 proteins have distant homologs outside the *Picornaviridae* family, indicating that external, yet-to-be identified sources were used as templates to incorporate ancestors of these three proteins into picornaviruses. While this interpretation may not be substantiated any further at the moment, the entire scenario of using a cellular sequence as a source of an innovation for a picornavirus finds strong support in an experimental study. Analysis of revertants of a mutant of poliovirus with severely decreased efficiency of proteolytic processing at the 3C-3D cleavage site identified a viable isolate carrying a short segment of rRNA incorporated in the viral genome in the vicinity of the original mutation[69]. This normally noncoding RNA encoded an in-frame amino acid sequence that apparently suppressed the effect of the original mutation, indicating positive selection as a driving force for the insertion to be fixed.

Some picornavirus species appear to have also been able to expand their proteome without using external templates for new genes. It was proposed that 2A1 evolved by duplication of the most conserved 3C protein in the ancestor of enteroviruses[46,238]. Both these proteins are prototypes of the only two known lineages of unique cysteine proteinases with a chymotrypsin-like fold[10,362]. Sequence and tertiary structure similarities between 2A1 and 3C are rather remote. Accordingly, highly conserved 2A1 proteins of the enterovirus genus lack two of six β-strands in the N-terminal β-barrel that are otherwise conserved in 3C and other structurally related proteases[362,430]. These and other differences between 2A1 and 3C proteases, including emergence of a unique Zn-binding site in 2A1[362,488], could be attributed to extensive divergent evolution following a duplication event. This interpretation is

now supported by the discovery of sapeloviruses[274,341,460] (see also [264]), forming a sister phylogenetic group to enteroviruses and encoding 2A1 proteins comparable with 3C proteases in size and lacking a Zn-binding site (Gorbalenya, unpublished) (Fig. 1 and 2).

Likewise, duplication was implicated in generating three tandem copies of 3B in FMDV, an aphthovirus[141], and two copies of 3B in the newly described SealPV, a phocavirus[250] (Gorbalenya, unpublished). These viruses are separated by a relatively large evolutionary distance, populated with several picornaviruses employing a single VPg (Fig. 1 and 2). The most parsimonious explanation for this phylogenetic pattern is that VPg duplication must have happened independently in the two picornavirus lineages, thus representing a rare case of parallel protein evolution. It is worth mentioning that the only currently known erbovirus, equine rhinitis virus B (ERBV; originally called ERV-2), encodes a 3B that is flanked from the N and C termini by sequences remotely resembling 3B[484] (Gorbalenya, unpublished). The upstream sequence was even coined a "pseudo-VPg"[484], and both may be remnants of the original 3B duplications that remained fused with the flanking proteins, 3A and 3C. If this were the case, the evolutionary history of 3B (VPg) triplication would be more complex than if it was restricted only to the FMDV species. In particular, the 3B triplication may have occurred either independently in the ERBV and FMDV lineages or in a more recent common ancestor of these viruses. In the latter scenario, the fate of the three 3Bs must be very different in the three descending species, ERBV, ERAV, and FMDV, that form a phylogenetically compact cluster (Fig. 1). Two of the three VPgs must have been either lost or deteriorated in ERAV and ERBV, while all three copies of VPg are present in FMDV. In this context, it is also interesting that dicistroviruses, a family of insect viruses resembling picornaviruses, include viruses that may employ different numbers of VPg[330].

The origin of the N-terminal amphipathic helix of 2C is another case open to different evolutionary interpretations. This helix is apparently conserved across all picornaviruses, yet its variety encoded by enteroviruses is most similar to the N-terminal amphipathic α-helix of NS5a protein, unrelated to 2C, that is encoded by HCV of the *Flaviviridae* family[447]. Interestingly, the HCV NS5a helix sequence affinity to the enterovirus 2C α-helix is comparable to that of its ortholog in NS5a of most related pestiviruses. This unusual pattern of conservation indicates that most similar varieties of α-helices of 2C and NS5A operate under common evolutionary constraints, while they are fused to unrelated proteins. These α-helices could have emerged independently in two families, being a case of convergent evolution, or they could be paralogs that emerged from a common ancestor, one by descent and another by nonhomologous recombination[447].

Besides expansion, gene loss could have also contributed to generating the proteome diversity in picornaviruses. Gene loss along with repeated introduction of a protein variety may be invoked for explaining phylogenetic discontinuity of the presence of the protein variety in picornaviruses. For instance, in the monophyletic enteroviruses and sapeloviruses (Fig. 1), avian sapelovirus is the only species among 13 identified so far that

has no 2A1 protein (or other 2A variant) encoded (Gorbalenya, unpublished). The most parsimonious explanation for this anomaly is that 2A1 was lost in an ancestor of this lineage through nonhomologous recombination. More complex scenarios must be drawn to accommodate the evolution of 2A4[306] and 2A6 proteins, each confined to two overlapping and phylogenetically separate subsets of picornaviruses (Fig. 1). One of these subsets comprises a phylogenetically compact cluster that is basal in the picornavirus tree and formed by four viruses of three genera, phoca- , avihepato-, and parechoviruses. Only one virus in this cluster, SealPV of phocaviruses and HPeV of parechoviruses, lacks either 2A6 or 2A4, respectively, indicative of a protein loss. In the sister lineage of this virus cluster that is formed by the remaining 10 genera, these proteins may have been lost repeatedly (Lauber and Gorbalenya, unpublished).

A distant evolutionary perspective on the picornavirus proteome

Above, we discussed the composition and evolutionary dynamic of the picornavirus proteome. With a large subset of proteins present in all picornaviruses, could we say, from a protein perspective, that they define what picornaviruses are?

To answer this question, the picornavirus proteome must be scrutinized in a broader evolutionary context of diverse viruses, collectively known as picorna-like viruses, which resemble picornaviruses in more than one aspect[20,154,160,270]. Over the last decade the number of these viruses has been steadily growing, and a subset of them most similar to picornaviruses is now taxonomically recognized as the *Picornavirales* order[287]. The latter is composed of seven families, including picornaviruses, and a number of unclassified viruses. As its name suggests, the order was coined after the picornaviruses, which were considered prototypic in the order. For the sake of this review it is important to note that all these viruses have proteomes that include counterparts for all ubiquitous picornavirus proteins and, in a few viruses, also a 2A variant (Fig. 5). These viruses are extremely diverse, and many have a genomic organization that differs from that of picornaviruses. For instance, clusters of capsid and replicative proteins, known as the capsid and replicative modules, respectively, are found on separate RNAs in most viruses of the plant *Secoviridae* family or in the permuted order relative to that of picornaviruses in the insect *Dicistroviridae* and unicellular organism *Marnaviridae* families[20,154,269,279,444]. On the other hand, the protein backbones of polyproteins of viruses in the families of *Picornaviridae* and *Iflaviridae*[158,231,280,344,400] and *Sequivirus* and *Waikavirus* genera of the *Secoviridae* family[383,384] are colinear. Viruses of the latter two virus families infect invertebrates and plants, respectively.

Because of the genomic colinearity and host range specifics, these families might be considered “picornaviruses” of invertebrates and plants, respectively. Practically, these genomic colinearities indicate that none of the proteins may serve as a molecular marker of the *Picornaviridae* family. To discriminate picornaviruses from iflaviruses and sequiviruses at

the sequence level, one needs to examine protein motifs. Two signatures can be noted in this respect. First, the picornavirus 3B is distinguished from VPgs of the ifla- and sequiviruses through the usage of Tyr versus Ser for nucleotidylation. Second, VP4 (when it is produced) is located upstream of VP2 and VP3 in picornaviruses and iflaviruses, respectively[300]. A more systematic comparative analysis might identify additional sequence characteristics specific to each of these three families.

The data discussed above indicate that the picornavirus genetic plan, with its polyprotein domain backbone, was born (long) before the emergence of the first ancestral picornavirus. It is not known whether the picornavirus genetic plan was the ancestral one or if it was derived from one of two closely related plans with either a permuted order of capsid and replicative modules or a bisegmented virus. The physical separation of the capside and replicative protein modules in bisegmented viruses of two picornavirus-like lineages implies that the coupling of these modules could be constrained less than that in picornaviruses. In this respect, it could be relevant that the polyprotein of caliciviruses includes a picornavirus-like replicative module that is fused with the capsid protein, which includes only one full copy of the jelly roll domain[189]. This simpler capside module's organization may have predated that of picornaviruses, for which expansion by a jelly roll domain triplication would be a relatively recent event. Consequently, the replicative module domain organization must be older than that of the capside module. These types of scenarios, involving comparative genomics of very distant virus families, can be tested in phylogenetic inference analyses, in which the direction of evolution is independently defined, a formidable challenge in virus evolution research. The scenarios could be also provide insights for back-rolling the early evolution of the replicative module of picornaviruses. Although these exercises may seem to be of remote relevance to understanding picornavirus proteome evolution, a link could be there. Some 25 years ago, several types of imperfect tandem repeats of several sizes with a common denominator of 11 running across the 2C-3D protein region of poliovirus polyprotein were uncovered[167,172]. They were interpreted as vestiges of a primordial multistep amplification process that gave rise to the proteins of the 2C-3D region. This scenario implies a concerted evolution of the three major ancient proteins, 2C, 3C, and 3D, of the replicative module, starting from the primordial stage of life. In this framework, picornaviruses may be seen as direct descendants of the ancestral self-replicating module. If this link between entities separated by a huge evolutionary distance is real, then it must manifest in functioning proteins as it does in the genome text. Verifying this link was beyond the realm of possibilities when the periodicity was discovered. Let's hope that future technical advancements and our understanding of the picornavirus proteome will make this testing approachable.

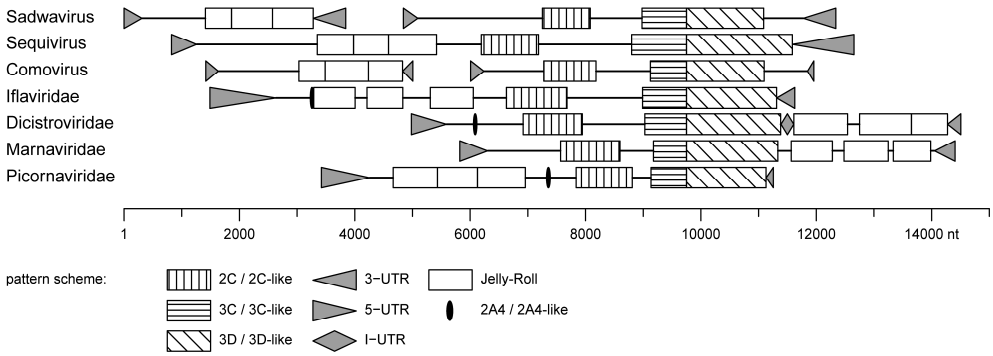


Figure 5. Conservation and diversity of genetic plans of the order *Picornavirales*. Genomic organizations for seven viruses are shown (similar to Fig. 2) and represent polyprotein layouts of families of the *Picornavirales*. Different shapes and shades were used to highlight protein families found in all or several virus families. Borders of proteins were identified using the GenBank annotation where available. Otherwise, positions were estimated utilizing homology searches (HMMer) against profiles of the picornavirus proteins[125]. The viruses included are Strawberry latent ringspot virus (Sadwavirus), Maize chlorotic dwarf virus (Sequivirus), Patchouli mild mosaic virus (Comovirus), Deformed wing virus (Iflavirus), Kashmir bee virus (Dicistrovirus), Heterosigma akashiwo RNA virus (Marnavirus), and encephalomyocarditis virus (Picornavirus). For Sadwavirus and Comovirus the two RNA segments are shown.

Concluding remarks

Since the time of a previous review on the evolution of the picornavirus proteome[192], considerable advancements have been made with sampling of the picornavirus genomic space. This multiteam endeavor has tripled the number of known picornavirus species and doubled the number of genera. Analysis of the emerged, and still growing, large body of information has confirmed previously uncovered trends and generated new insights into protein evolution. Chiefly, the interplay between the concerted evolution of the backbone, mostly ancient, genes descending from the ancestral picornavirus and the modular evolution of L and 2A genes is evident in old and new picornavirus lineages. A previously identified diverse repertoire of proteins encoded by the L and 2A loci keeps steadily expanding, revealing new functional partners of the most conserved picornavirus proteins. Gene duplication, overprinting, loss, and horizontal acquisition from cellular and viral genes were reaffirmed to shape the picornavirus proteome, likely through nonhomologous recombination. It is also now apparent that parallel (convergent) emergence of a protein (3B) may have contributed to picornavirus evolution. Also, it was recognized for the first time that the 2A locus could be multicistronic, encoding two or three proteins.

The fast-accumulating knowledge about the genomic diversity of picornaviruses is most critical for current and future analytical efforts. We may expect to see a shift from cataloging the protein diversity to comprehending molecular details of how the proteome has evolved and what forces and restraints operated behind and operate now in different

lineages. For instance, we know surprisingly little about the identities of genes that served as sources of major innovation in the diversification of the picornavirus proteome. Identifying these sources could provide insights also for understanding the molecular environment in which picornaviruses replicate and the modus operandi of the replicase machinery. Protein evolution must also be considered in the context of genomic evolution, including its RNA signals in the coding and noncoding regions. Indeed, picornaviruses have accommodated large-scale evolution in the UTRs[205] and cis-acting RNA elements in the ORF[164,486]. It should not come as a surprise if evolution of the picornavirus RNA and proteome turns out to be coupled. The emergence of diverse ORFans in the L region of the genome may be an example of this dynamic relationship.

Although picornaviruses are commonly considered to be vertebrate viruses, the verified host range of picornaviruses is confined to a few mammals and birds. Extending the virus discovery effort to other vertebrate species should fill a huge gap in our knowledge and bring multifold benefits, including improved understanding of protein evolution. We could learn how the host constrains proteome diversity and its evolution and how the proteome evolutionary dynamics is shaped in different lineages. The latter should include the entire diversity range from species to genera to families and to orders. With comprehensive host coverage, the contemporary protein universe of picornaviruses may be revealed. This could (and should) facilitate reconstructing its past, in time and space, and predicting future trends, both of which would contribute to our understanding of the fundamentals of picornaviruses. These exercises could also be equally insightful for developing innovative strategies to control picornavirus infections, defining targets for antivirus drugs[82] (see [93]), and improving designs of picornavirus-based vectors[16,493].

Acknowledgments

We are indebted to Alexander Kravchenko, Dmitry Samborsky, and Igor Sidorov for administration of the Viralis software platform that was used in preparing figures for the chapter. Work in the A. E. Gorbalenya group was partially supported by The Netherlands Bioinformatics Center, EU FP6, Collaborative Agreement in Bioinformatics, between LUMC and MSU, and by the Leiden University Fund.

CHAPTER 8

General Discussion

Uncovering barriers to genetic divergence of RNA viruses

When searching the public sequence database GenBank/RefSeq[36] using the keyword 'virus' roughly 1.8 million nucleotide entries (April 2012) show up. Perhaps not surprisingly, a large fraction is contributed by human immunodeficiency virus 1, influenza A virus and hepatitis C virus due to their high medical relevance and outstanding efforts to study theses pathogens. However, this wealth of genetic data also involves a great many other viruses both known and unknown ones, the latter mostly represented by so-called genomic survey sequences from metagenomics studies. On the one hand this development allows us to study the genetic diversity and evolution of viruses in unprecedented detail, but, on the other hand, it increasingly challenges virus taxonomy since other types of virus characterization normally cannot keep up with the pace set by genome sequencing. Hence, it is tempting to consider genome sequences as the ultimate source of information to be utilized in virus taxonomy, a notion which is supported by additional facts. First, they bring highly accurate knowledge due to the low sequencing error rates which generally fall well below 1%[476]. Second, sequencing is fast and relatively cheap nowadays[278] as evermore efficient techniques are being developed[395]. Third, genome sequences are easy to digitize and compare, readily enabling quantitative analyses. And last but not least, the use of nucleotide sequences as carriers of genetic information and heredity presents a universal property common to all biological entities on earth.

In this thesis, a computational approach (DEmARC) for virus classification basing solely on genome sequences was developed (see chapter 2). It involves the calculation of genetic distances between all pairs of viruses considered and the partitioning of the resulting distance distribution to delimit both the levels and the taxa of a hierarchical classification. This methodology is, however, not new to virus taxonomy and can be traced back as far as 1988 when Shukla and Ward conducted a groundbreaking study in which they classified various potyvirus strains using only coat protein sequences[422]. The approach was adapted later in taxonomic studies to classify viruses of various families and genome types including picornaviruses[67,339], caliciviruses[413,496], geminiviruses[142], coronaviruses[90,162], potyviruses[2], fexiviruses[1], papillomaviruses[38,94], poxviruses[290], rotaviruses[317], and hantaviruses[309]. Furthermore, the use of divergence thresholds of genes or sets of genes for assisting virus classification is becoming common practice in virus taxonomy[257]. There is, however, no consensus on key parameters of the method among the different studies, which includes (i) what genome regions to include, (ii) what sequence type – nt or aa – to use, (iii) what type of alignments – pairwise or multiple – to compile, and (iv) what measure of sequence similarity – uncorrected percentage identities or distances that correct for multiple substitutions at the same site – to calculate. DEmARC-based results from this thesis (see chapter 2) suggest that a classification basing on a multiple alignment of all proteins conserved across all viruses considered is most consistent and stable, although this finding needs to be verified for other virus families. Furthermore, the use of evolutionary-based corrected distances,

despite having little impact on low levels (e.g. species), should be favored over uncorrected distances since the latter show poor resolution at higher levels (e.g. subfamily) due to the accumulation of multiple substitutions at the same sequence site[117,121]. A similar trend was observed in the DEmARC-based analysis of corona- and toro-/bafiniviruses (Lauber and Gorbalenya, in preparation). The resulting classification already formed the basis to revise the taxonomy of the family *Coronaviridae*, which involved the introduction of two subfamilies (formed by coronaviruses and toro-/bafiniviruses, respectively), the recognition of five genera (formed by group 1, group 2, group 3 coronaviruses, toroviruses, and bafiniviruses, respectively), and the revision of several coronavirus species accompanied with an expansion of the host range[90].

Recently, the PASC tool[24], promoted by NCBI, has emerged as the standard for pairwise-distance-based virus classification and was utilized in several of the studies mentioned above. A main objective of PASC is the speedy classification of newly identified viruses with sequenced genomes. To do so, PASC depends on pre-established classifications, usually brought by the ICTV taxonomy, for roughly 50 families or floating genera. A new virus is classified using thresholds on its similarity to taxa in the respective pre-established classification. Importantly, these demarcation thresholds have been defined *a priori* as the lowest intra-level (for instance intra-species) similarity observed across virus pairs in the pre-established classification. Exactly this approach, however, presents a potential pitfall of the method since no valid golden standard classification is available for any virus family, due to the intrinsic lack of fossil data in virology. Hence, PASC is suitable for initial classification of newly identified viruses if an advanced taxonomy of the respective family/genus is available, but may produce unreliable results in other cases. Only few studies tried to approach the problem of threshold determination objectively, not depending of any pre-existing classification. Matthijnssens et al.[317], for example, selected the threshold at which the ratio of intergroup to intragroup sequence identities dropped below one. Unfortunately, the basis for this choice was neither explained in detail nor evaluated rigorously. As shown in this thesis, DEmARC enables the user to measure the support for both the demarcation thresholds and the inferred taxa in a quantitative manner, thus allowing for an objective selection of thresholds and resulting virus groupings. This discriminates the approach from any previous study in virus taxonomy.

Importantly, DEmARC can serve not only pure classification purposes but also the prediction of biological properties of the analyzed viruses and the inferred virus taxa (see chapter 3). Intuitively, a newly identified virus classified as belonging to a known virus species is expected to show phenotypic properties similar to that of other viruses in the species, a prediction which is also available through traditional virus taxonomy. Yet, the predictive power of a DEmARC-based classification extends to non-traditional projections, due to the fact that virus taxa of the same level are delimited objectively by applying the same criterion (a demarcation threshold on genetic divergence) universally to all viruses at hand. One such prediction concerns the natural genetic diversity of a taxon, that might be

heavily underestimated by the current virus sampling, but which can be predicted by utilizing information from other, well-sampled taxa of the same level. Hence, DEmARC offers the means to identify those taxa on which to focus future virus discovery efforts in order to obtain a comprehensive picture of the natural genetic diversity of the virus family/genus under consideration. Notably, the predicted genetic diversity of a taxon presents only an upper limit of the actual natural diversity. For instance the moderately sampled picornavirus species *Hepatitis A virus*[144,263] (around 50 available complete genome sequences between 1993 and 2010) shows a relatively low genetic diversity (see Fig. 3 in chapter 3), which may be due to the unusually low evolutionary rate exhibited by these viruses[198,326,403]. From a more fundamental perspective, the presence of peaks and valleys in the pairwise distance distribution, like those seen for the well-sampled family *Picornaviridae*, may provide an insight into commonalities across viral lineages during evolution. Specifically, the observed distance discontinuities at and above the genus level and the distance peaks separated by these discontinuities could be explained by periods of, respectively, mass extinction and mass speciation of viral lineages, possibly reflecting large-scale changes in the environment that had a bearing on their hosts. Here, distance discontinuity is defined as a distance range with zero or marginal frequency below a certain noise level.

It should be noted that there is a long-lasting dispute on the use of pairwise distances as a single criterion for classifying viruses[467]. This dispute is largely linked to the question whether virus species, forming the basic level in virus taxonomy[373], are real biological entities[44,259,322] or simply constructs in our mind[466] developed for the convenience of biologists. If the former is true then certain biological properties, which could be used to discriminate between virus species, are expected to exist. In the case of eukaryotic organisms, for which it is generally accepted that species are evolving biological entities[86], such properties usually include genetic incompatibilities between species that result in separated gene pools. It is tempting to apply this biological species concept, originally introduced by Dobzhansky in 1937 [104], to viruses, although its validity is questioned for organisms that reproduce asexually[207]. Nevertheless, as shown in this thesis, virus clusters (of the lowest level) can be delimited genetically at the family level (see chapter 3) or even across related families (see chapter 4) through distance discontinuity in the conserved proteins. This distance discontinuity, which is the result of inter-virus distances being generally lower inside a cluster than between clusters, is nontrivial and could be explained by only two causes (when assuming that the calculated distances adequately estimate the real genetic distances; technical causes, like certain bias in the estimation of pairwise distances, are most likely not the reason for the observed distance discontinuity because these causes would be expected to mask rather than produce such signals). First, the distance discontinuity could be due to insufficient sampling of both the number and the diversity of the analyzed sequences. If this is not the case, as presumably for the family *Picornaviridae* with its numerous species distributed over a dozen or so genera[263] and sampled by more than 1200 sequences, the observed distance discontinuity is likely due to

biological factors enforcing constraints that limit the divergence of viruses of the same but not of different clusters. In the latter context, it could be argued that the delineated virus clusters correspond to biological species. A plausible factor of speciation could be the action of homologous recombination if restricted to viruses of the same species, and this was already suggested for picornaviruses[305]. In this way, homologous recombination could resemble the exchange of genetic material during (ordinary) sexual reproduction, thereby setting a barrier to genetic divergence and, consequently, to speciation, whereas viruses from different species continue to diverge by mutation. If true, not only would it provide a biological foundation for the recognition of virus species as real, evolving entities but also should it pin medical relevance to species[238,453]. This would have great implications for many branches of virology including virus diagnostics, antiviral research, and epidemiological studies.

Uncovering barriers to gene length in RNA viruses

Comparative sequence analysis in virology is usually concerned with genetic variation (nucleotide or amino acid differences among the compared sequences) and its utilization for making biological inferences like structural predictions, functional predictions for a sequence or specific sequence residues, or reconstructing the evolutionary history of the sequences. There is, however, a second dimension that receives little attention so far: the length of genetic sequences (in number of nt or aa). This includes both the total size of a viral genome and the size of genome regions encoding functional elements, for instance proteins. It is generally acknowledged that the genome size of RNA viruses is strongly constrained as a result of (i) the low fidelity of their polymerases[216] which would drive larger genomes into an 'error catastrophe'[61,134], (ii) the selection for high replication speed[33,135], and (iii) the relative inflexibility in expanding the virions of viruses with icosahedral capsids in order to accommodate larger genomes[76]. Hence, RNA virus genome sizes are in the range from two to 32 kb with an average of about 10 kb (Fig. 1). When counting only the size of the largest genome segment (single RNA molecule) nidoviruses with genomes above 20 kb, which comprise coronaviruses, toro-/bafiniviruses, roniviruses, and mesoniviruses, outcompete all other known RNA viruses. Still, and this also applies to nidoviruses, all genes of an RNA virus must be compressed into a confined genomic space. As a consequence, genome regions often show multiple functions, which is achieved, for instance, through overlapping ORFs[174,272] or the encoding of RNA regulatory elements inside a protein-coding gene[319]. As would be expected, ORF overlap was found to be largest for viruses with the smallest genomes and vice versa[34].

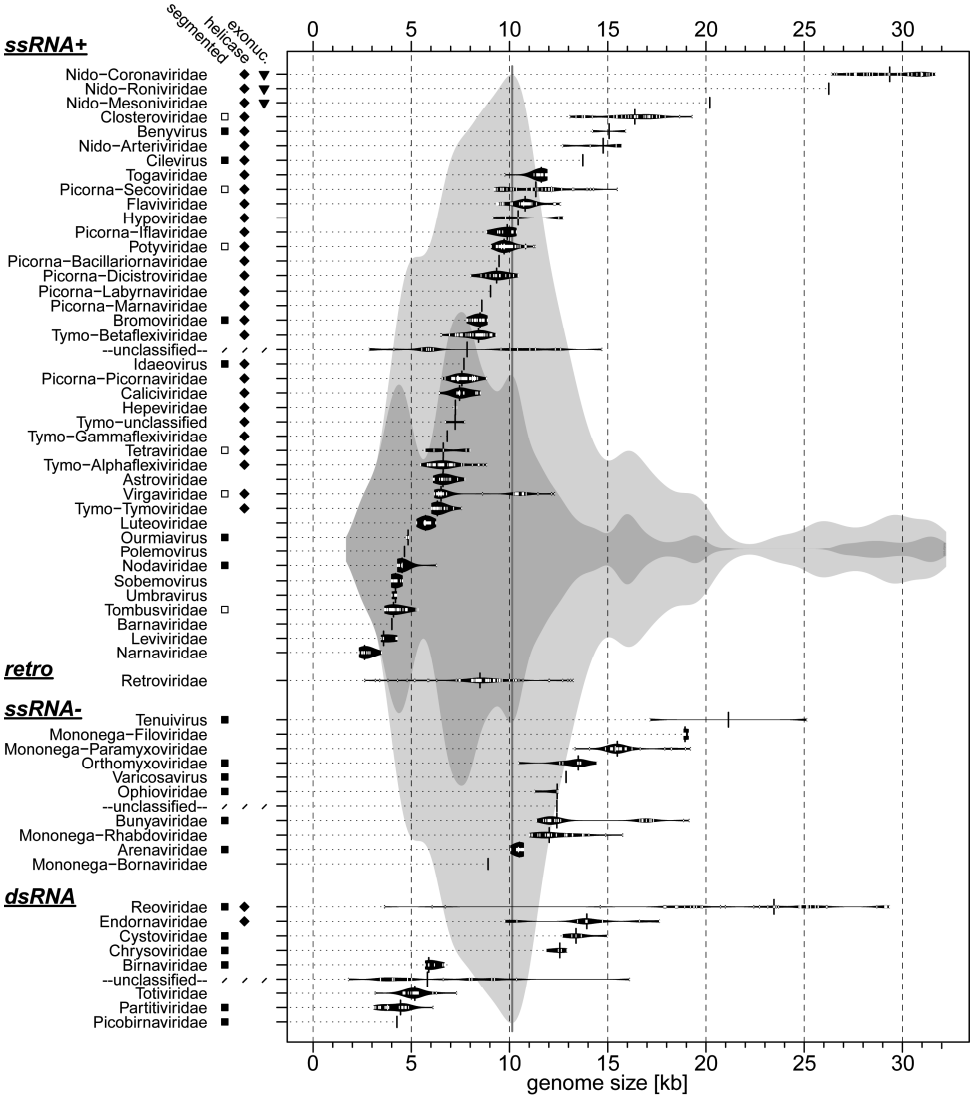


Figure 1. Genome sizes of RNA viruses and the relation to genome segmentation, and helicase and exoribonuclease expression. Shown are beanplots[247] of genome sizes for all known families or floating genera of viruses with ssRNA+, ssRNA-, and dsRNA genomes and retroviruses. Sizes were extracted from the Viral Genomes Resource at NCBI[25] (March 2012). A bean (black shape) shows the density distribution of individual genome sizes (white vertical bars) for a virus group. The median genome size per group is indicated by black vertical bars and was used for sorting. Additional information is shown next to virus group names: some (open square) or all (filled square) viruses have segmented genomes; expression of a helicase (diamond); expression of an exoribonuclease (triangle); / not applicable. Note that retroviruses don't encode a helicase due to specifics in their replication cycle[244]. The joint distribution of all RNA virus genome sizes and all sizes of the largest genome segment are shown, respectively, as a light-gray and a dark-gray bean in the background. The average RNA virus genome size is indicated by the gray vertical line. Ideas adapted from [174] and [175].

The exceptionally large genomes of nidoviruses, which exceed about threefold the average size of a ssRNA+ genome, uniquely encode a 3'-5' exoribonuclease (Fig. 1). It was proposed that this enzyme improves the otherwise low fidelity of RNA virus replication[432], and allowed a subset of nidoviruses to overcome a genome size threshold of about 20 kb (see chapter 5). Furthermore, a second threshold of about 8 kb set by astroviruses seems to be associated with the expression of a protein with NTP-binding motif indicative of a helicase (Fig. 1). This strong correlation was already noticed before[176] and it was proposed that the helicase acquisition allowed the enlargement of RNA virus genomes above the observed threshold[174]. The helicase expression seems to also be associated with the size of the largest genome segment of dsRNA viruses but not with retrovirus and ssRNA- virus genomes (Fig. 1). The latter observation indicates that the unwinding of long stretches of duplex RNA (for instance during replication and transcription), which is the expected primary function of a viral helicase, could be achieved through unconventional mechanisms[244], or, alternatively, the formation of such dsRNA structures would need to be prevented in the first place. Notably, other protein domains, besides the exoribonuclease and helicase, do not show a strong association with genome size in RNA viruses[175].

Yet, theses findings show, at least in the case of dsRNA and ssRNA+ viruses, that the expression of RNA-processing enzymes with specific functions allows some RNA viruses to employ larger genomes than those that lack these enzymes, which results in a considerable variety of genomic sizes. This is contrary to the prevailing perception that RNA viruses are simply limited in genome size with essentially no variability (concerning both the whole genome and distinct genetic elements within the genome) which would be worth a detailed analysis. In fact, different genome regions (e.g. genes) might be constrained differently depending on the encoded function. It was shown, for instance, that the gene length of polymerases, but not of nucleocapsids, increases with increasing genome size among RNA viruses, and the authors link this size increase of the protein to an improved replication fidelity[33]. Beyond these large-scale analyses, however, little research on specifics about the regulation of gene length during RNA virus evolution is available.

In this thesis, size constraints on genes or gene sets were analyzed at the scale of a virus family or multiple related families. In the case of the family *Picornaviridae* a striking negative correlation between sequence conservation and size variation of viral proteins was observed. Specifically, the sizes of the six proteins conserved in sequence across picornaviruses (three capsid proteins plus helicase, proteinase, and polymerase) proved to be constrained most strongly (see chapter 7). The polymerase, for instance, shows the highest sequence conservation and varies in size among known picornaviruses by not more than 40 aa, which is about 8% of its total 460 aa, with a dispersion of only eight aa (data not shown). On the one hand, this is not surprising since the most conserved proteins are essential to the virus and, thus, their function must be retained during evolution. These six picornavirus proteins control the three main steps in the virus life cycle – genome replication, genome expression, and encapsidation. On the other hand, the apparent barrier to size

variation of the conserved proteins is nontrivial given that these proteins accepted as much as three replacements per residue (on average) when comparing the most distant picornavirus pair (see chapter 2). This shows that evolutionary constraints on picornavirus proteins act not only in the dimension of aa substitutions but also in a second one, the size of these proteins, by limiting the amount of insertions and deletions. These fundamental findings were supported by a similar analysis of nidoviruses, in which the sizes of key replicative enzymes (those encoded in ORF1b) within a family were found to be most strongly constrained as well (see chapter 6). Moreover, the combined set of these ORF1b-encoded proteins expanded first in the transition from small nidovirus genomes of at most 16 kb (resembled by contemporary arteriviruses) to large genomes of at least 20 kb (mesoni-, roni-, toro-, and coronaviruses). Specifically, the abovementioned exonuclease and two methyltransferases have been inserted in ORF1b[73,432]. This supports the dominant role of replicative proteins in the control of gene/genome length during RNA virus evolution. However, a second stage of expansion (from 20 to 26 kb) predominantly involved ORF1a, indicating that another, still illusive factor was acquired that allowed some nidoviruses to expand their genomes even further. Only after this second stage, the third main genome region – the 3'-proximal ORFs (3'ORFs) – expanded through the acquisition of genes with diverse, often unknown functions that may vary even between closely related nidoviruses[174]. In summary, these findings suggest a functional hierarchy of the three genome regions (ORF1b, ORF1a, and 3'ORFs) in the control of gene and genome size during evolution. Importantly, the three regions are characterized not only by different expression mechanisms which results in unequal molecular amounts of protein products, but also by different degrees of genetic divergence (proteins encoded in ORF1b and 3'ORFs show the highest and lowest sequence conservation, respectively). It should be noted that this hierarchical model defines universal constraints that have acted independently and simultaneously on each nidovirus lineage during evolution. Contemporary nidoviruses may have reached different points on the trajectory of genome expansion. Arteriviruses, for example, seem to be unable to overcome a barrier to genome size at around 16 kb due to missing factors in ORF1b that includes the exonuclease, whereas mesoniviruses, frozen in a stage of intermediate genome size of around 20 kb, are lacking a different factor predicted to be located in ORF1a.

Besides bringing important fundamental insights, the relationship of sequence conservation and size variation has immediate practical implications. For instance, it could provide guidance for key decisions in genetic engineering experiments, as it predicts where the insertion of the gene of interest will likely compromise the virus (namely in ORF1b) and where it will not (3'ORFs and, possibly, ORF1a). Nidoviruses at different points of the genome expansion trajectory may differ in this respect. An analogous reasoning can be applied to the L and 2A regions of the picornavirus genome, which show a large diversity of encoded proteins and, consequently, the largest tolerance of size variation (see chapter 7). Moreover, this relation and the resulting practical implications can be extended to the level

of a single gene/protein, as shown in a recent study of poliovirus, where it was found that the insertion of short nucleotide stretches is only tolerated at gene regions of low sequence conservation[448]. These insertions could be used for tagging specific proteins in order to purify the protein or associated protein complexes from infected cells, or to visualize the location and dynamics of a protein over time.

Future prospects

Genetics-based classification by DEmARC offers the means to modeling the evolution of the genetic diversity of the viruses under consideration. Like any model that approximates nature, it is limited by certain simplifications. One such simplification is presented by the substitution model used to estimate pairwise genetic distances. For protein sequences it is common practice to utilize aa substitution rate matrices that are pre-compiled empirically on specific protein training sets. Because of the extreme mutation rates of RNA viruses, the WAG substitution matrix was preferred in this thesis, since it is (i) estimated by ML and (ii) based on a large variety of protein families[478]. However, it may be worth exploring the impact of mores sophisticated matrices, possibly trimmed to RNA virus sequences, if available in the future. Such a matrix already brought valuable insight into the evolution of reverse transcriptases in retroviruses, but was highly specific to the analysis of this particular protein[102]. Furthermore, allowing for heterogeneity of the substitution rate across viral lineages, which is currently prohibited in practice by its high computational costs, may fit the evolution of the viruses at hand more adequately, especially when certain lineages show an elevated rate, like suggested for human rhinoviruses[318,427]. Another data-related aspect, the impact of which could be explored, is the additional sequence variation that is accumulated during propagating the viruses in cell culture before sequencing. This concerns perhaps the majority of viral sequences from public databases, but its scale is expected to be limited and, thus, should not have an effect on the demarcation threshold of the species or higher levels. Future research efforts should also involve scrutinizing other classification approaches that rely on genetic sequences. Among them are phylogeny-based techniques like the branching index[223], which can be used to infer statistically whether a query sequence clusters with a known clade in the tree, and a method that determines an increase of the branching rate in the tree to define the species boundary[368]. The latter study is from the field of *DNA barcoding*, a recently emerged line of research that aims at genetics-based taxonomy of cellular organisms[64,204]. Most of these methods[204,289,379,406] face similar limitations like their counterparts in virus classification, including the dependency on a golden standard[320], as well as additional challenges owing to the large sizes of cellular genomes[64]. Nevertheless, genetics-based RNA virus classification should not continue to ignore such parallel developments. Finally, future studies should be devoted to the analysis of viruses from other families/orders including those that don't have ssRNA+ genomes, and there are promising preliminary results for mononegaviruses, which have ssRNA- genomes

(Lauber and Gorbalenya, data not shown and [281]). Such large-scale analyses could help not only to validate the DEmARC approach but also to address important biological questions and, possibly, to reveal commonalities between virus families or orders. It could be asked, for instance, whether the DEmARC approach can be generally adopted for the recognition of genotypes of a virus species, much like it was done for human entero- and rhinoviruses[339,427], and whether these genotypes correspond to serotypes[277,288,407]. Further research in this direction is under way (Gorbalenya, personal communication). Moreover, the time of emergence of virus species from different families could be estimated using state-of-the-art tools[119], and analyzed in the context of fossil records of their hosts, in order to understand the dynamics of alternative processes like virus-host cospeciation and crossing of the host species barrier. What's more, the hypothesis that virus species are real biological entities that maintain a common gene pool by means of homologous recombination could be probed experimentally[238]. If it turns out to be true, it can only be explained by bringing a selective advantage for a virus species. Recently, it was shown that coevolution with a bacterial pathogen selects for sexual reproduction in *Caenorhabditis elegans* (which can also reproduce through self-fertilization)[327], and it could be speculated that a similar reasoning can be applied to the pathogen, in which case sexual reproduction would be defined more generally as the exchange of genetic material during replication, for instance by homologous recombination in the case of virus species. Moreover, the applicability of the biological species concept to asexual organisms is further supported by a recent study of stains of an archaeon that were found to form two persistently coexisting groups that exhibit high levels of homologous gene flow within each group and decreasing rates between groups in nature, indicative of ongoing sympatric speciation[62].

A proper classification of the viruses of interest is a prerequisite for many studies in virology by providing the units for which to measure the desired properties. This includes the analysis of sequence divergence and size variation of genetic elements, the second major topic of this thesis, in which the units were formed by species or genera of the same or closely related families. Future studies should be devoted to the analysis of additional families in order to verify the observed correlation (that genome size differences between relatively closely related viruses are the result of expanding or shrinking genomic regions poorly conserved in sequence) beyond picorna- and nidoviruses. Moreover, further insights into the emergence of the largest known RNA genomes employed by nidoviruses might be gained by including viruses not belonging to the order. In this respect it would be natural to consider barna-, sobemo-, luteo- and astroviruses[174] since they show the same genomic organization (ORF1a, ORF1b, 3'ORFs) but have much smaller genomes than nidoviruses (Fig. 1). Equally important would be a broader coverage of the natural diversity within the order *Nidovirales*, especially in the genome size range between arteri- and roniviruses which is currently represented only by the two mesoniviruses. This may also help to determine the additional factor(s), predicted to be located in ORF1a, which allowed roni-, toro- and coronaviruses to expand their genomes beyond that of mesoniviruses. Such analyses could

finally help to determine the ultimate upper limit of the RNA (virus) genome size, which, in turn, would contribute to our understanding of fundamental evolutionary processes like the proposed transition from RNA- to DNA-based life forms[153,159].

References

1. **Adams MJ, Antoniw JF, Bar-Joseph M, Brunt AA, Candresse T, Foster GD, Martelli GP, Milne RG, Fauquet CM**. (2004) The new plant virus family Flexiviridae and assessment of molecular criteria for species demarcation. *Archives of Virology* 149(5):1045.
2. **Adams MJ, Antoniw JF, Fauquet CM**. (2005) Molecular criteria for genus and species discrimination within the family Potyviridae. *Archives of Virology* 150(3):459.
3. **Adams P, Kandiah E, Effantin G, Steven AC, Ehrenfeld E**. (2009) Poliovirus 2C Protein Forms Homo-oligomeric Structures Required for ATPase Activity. *Journal of Biological Chemistry* 284(33):22012.
4. **Agol VI**. (2002) Picornavirus genetics: an overview. Semler,B., Wimmer,E. Molecular Biology of Picornaviruses. 269. Washington, DC, *ASM Press*.
5. **Agol VI**. (2002) Picornavirus genome: an overview. Semler,B., Wimmer,E. Molecular Biology of Picornaviruses. 127. Washington, DC, *ASM Press*.
6. **Agol VI**. (2010) Picornaviruses as a Model for Studying the Nature of RNA Recombination. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 239. *ASM Press*.
7. **Agol VI, Gmyl AP**. (2010) Viral security proteins: counteracting host defences. *Nature Reviews Microbiology* 8(12):867.
8. **Agol VI, Paul AV, Wimmer E**. (1999) Paradoxes of the replication of picornaviral genomes. *Virus Research* 62(2):129.
9. **Ahola T, Laakkonen P, Vihinen H, Kaariainen L**. (1997) Critical residues of Semliki Forest virus RNA capping enzyme involved in methyltransferase and guanylyltransferase-like activities. *Journal of Virology* 71(1):392.
10. **Allaire M, Chernaia MM, Malcolm BA, James MNG**. (1994) Picornaviral 3C Cysteine Proteinases Have A Fold Similar to Chymotrypsin-Like Serine Proteinases. *Nature* 369(6475):72.
11. **Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ**. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389.
12. **Aminev AG, Amineva SP, Palmenberg AC**. (2003) Encephalomyocarditis virus (EMCV) proteins 2A and 3BCD localize to nuclei and inhibit cellular mRNA transcription but not rRNA transcription. *Virus Research* 95(1-2):59.
13. **Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, Hilgenfeld R**. (2002) Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *Embo Journal* 21(13):3213.
14. **Anantharaman V, Aravind L**. (2003) Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biology* 4(2).
15. **Andino R, Rieckhof GE, Baltimore D**. (1990) A Functional Ribonucleoprotein Complex Forms Around the 5' End of Poliovirus Rna. *Cell* 63(2):369.
16. **Andino R, Silvera D, Suggett SD, Achacoso PL, Miller CJ, Baltimore D, Feinberg MB**. (1994) Engineering Poliovirus As A Vaccine Vector for the Expression of Diverse Antigens. *Science* 265(5177):1448.
17. **Ando T, Noel JS, Fankhauser RL**. (2000) Genetic classification of "Norwalk-like viruses". *Journal of Infectious Diseases* 181:S336-S348.

18. **Antonov IV, Leontovich AM, Gorbalenya AE** (2008) BAGG (Blocks Accepting Gaps Generator); http://www.genebee.msu.su/~antonov/bagg/cgi/bagg.cgi
19. **Arden KE, Mackay IM**. (2010) Newly identified human rhinoviruses: molecular methods heat up the cold viruses. *Reviews in Medical Virology* 20(3):156.
20. **Argos P, Kamer G, Nicklin MJH, Wimmer E**. (1984) Similarity in Gene Organization and Homology Between Proteins of Animal Picornaviruses and A Plant Comovirus Suggest Common Ancestry of These Virus Families. *Nucleic Acids Research* 12(18):7251.
21. **Arita M, Zhu SL, Yoshida H, Yoneyama T, Miyamura T, Shimizu H**. (2005) A Sabin 3-derived poliovirus recombinant contained a sequence homologous with indigenous human enterovirus species C in the viral polymerase coding region. *Journal of Virology* 79(20):12650.
22. **Baliji S, Cammer SA, Sobral B, Baker SC**. (2009) Detection of Nonstructural Protein 6 in Murine Coronavirus-Infected Cells and Analysis of the Transmembrane Topology by Using Bioinformatics and Molecular Approaches. *Journal of Virology* 83(13):6957.
23. **Baltimore D**. (1971) Expression of Animal Virus Genomes. *Bacteriological Reviews* 35(3):235.
24. **Bao Y, Kapustin Y, Tatusova T**. (2008) Virus Classification by Pairwise Sequence Comparison (PASC). Mahy,B.W.J., Van Regenmortel,M.H.V. Encyclopedia of Virology, 5 vols. 342. Oxford, *Elsevier. Vol. 5*.
25. **Bao YM, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T**. (2004) National Center for Biotechnology Information Viral Genomes Project. *Journal of Virology* 78(14):7291.
26. **Baranowski E, Ruiz-Jarabo CM, Pariente N, Verdaguer N, Domingo E**. (2003) Evolution of cell recognition by viruses: A source of biological novelty with medical implications. *Advances in Virus Research, Vol 62* 62:19. Advances in Virus Research.
27. **Barrette-Ng IH, Ng KKS, Mark BL, van Aken D, Cherney MM, Garen C, Kolodenko Y, Gorbalenya AE, Snijder EJ, James MNG**. (2002) Structure of arterivirus nsp4 - The smallest chymotrypsin-like proteinase with an alpha/beta C-terminal extension and alternate conformations of the oxyanion hole. *Journal of Biological Chemistry* 277(42):39960.
28. **Bazan JF, Fletterick RJ**. (1988) Viral Cysteine Proteases Are Homologous to the Trypsin-Like Family of Serine Proteases - Structural and Functional Implications. *Proceedings of the National Academy of Sciences of the United States of America* 85(21):7872.
29. **Beckner M**. (1959) The biological way of thought. New York, *Columbia University Press*.
30. **Beijerinck MW**. (1898) Over een contagium vivum fluidum als oorzaak van de vlekziekte der tabaksbladen. *Versl.Gew.Verg.Wiss.en Natuurk.Afd., Kon.Aka.Wetensch.Amst.* 7:229.
31. **Belalov IS, Isaeva OV, Lukashev AN**. (2011) Recombination in hepatitis A virus: evidence for reproductive isolation of genotypes. *Journal of General Virology* 92:860.
32. **Belshaw R, de Oliveira T, Markowitz S, Rambaut A**. (2009) The RNA Virus Database. *Nucleic Acids Research* 37:D431-D435.
33. **Belshaw R, Gardner A, RarnbaUt A, Pybus OG**. (2008) Pacing a small cage: mutation and RNA viruses. *Trends in Ecology & Evolution* 23(4):188.

34. **Belshaw R, Pybus OG, Rambaut A**. (2007) The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research* 17(10):1496.
35. **Bennett SP, Lu L, Brutlag DL**. (2003) 3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence motifs. *Nucleic Acids Research* 31(13):3328.
36. **Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW**. (2010) GenBank. *Nucleic Acids Research* 38:D46-D51.
37. **Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL**. (2008) GenBank. *Nucleic Acids Research* 36:D25-D30.
38. **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(1):70.
39. **Bessaud M, Joffret ML, Holmblat B, Razafindratsimandresy R, Delpeyroux F**. (2011) Genetic Relationship between Cocirculating Human Enteroviruses Species C. *Plos One* 6(9).
40. **Bhardwaj K, Guarino L, Kao CC**. (2004) The severe acute respiratory syndrome coronavirus Nsp15 protein is an endoribonuclease that prefers manganese as a cofactor. *Journal of Virology* 78(22):12218.
41. **Bhardwaj K, Palaninathan S, Alcantara JMO, Yi LL, Guarino L, Sacchettini JC, Kao CC**. (2008) Structural and functional analyses of the severe acute respiratory syndrome coronavirus endoribonuclease Nsp15. *Journal of Biological Chemistry* 283(6):3655.
42. **Bhardwaj K, Sun JC, Holzenburg A, Guarino LA, Kao CC**. (2006) RNA recognition and cleavage by the SARS coronavirus endoribonuclease. *Journal of Molecular Biology* 361(2):243.
43. **Biebricher CK, Eigen M**. (2005) The error threshold. *Virus Research* 107(2):117.
44. **Bishop DHL**. (1985) The Genetic-Basis for Describing Viruses As Species. *Intervirology* 24(2):79.
45. **Blinkova O, Kapoor A, Victoria J, Jones M, Wolfe N, Naeem A, Shaukat S, Sharif S, Alam MM, Angez M et al.** (2009) Cardioviruses Are Genetically Diverse and Cause Common Enteric Infections in South Asian Children. *Journal of Virology* 83(9):4631.
46. **Blinov VM, Donchenko AP, Gorbalenia AE**. (1985) Internal Homology in the Primary Structure of Polio Virus Polyprotein - the Possible Existence of 2 Virus-Specific Proteinases. *Doklady Akademii Nauk Sssr* 281(4):984.
47. **Blinov VM, Gorbalenia AE, Donchenko AP**. (1984) The Structural Similarity Between Poliovirus Cysteine Proteinase P3-7C and Cellular Serine Proteinase of Trypsin. *Doklady Akademii Nauk Sssr* 279(2):502.
48. **Boonyakiat Y, Hughes PJ, Ghazi F, Stanway G**. (2001) Arginine-glycine-aspartic acid motif is critical for human parechovirus 1 entry. *Journal of Virology* 75(20):10000.
49. **Boursnell MEG, Brown TDK, Foulds IJ, Green PF, Tomley FM, Binns MM**. (1987) Completion of the Sequence of the Genome of the Coronavirus Avian Infectious-Bronchitis Virus. *Journal of General Virology* 68:57.
50. **Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M**. (2008) Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456(7224):942.
51. **Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E**. (2010) In Vitro Reconstitution of SARS-Coronavirus mRNA Cap Methylation. *Plos Pathogens* 6(4).

52. **Bouvet M, Imbert I, Subissi L, Gluasis L, Canard B, Decroly E**. (2012) RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proceedings of the National Academy of Sciences of the United States of America* 109(24):9372.
53. **Brian DA, Baric RS**. (2005) Coronavirus genome structure and replication. *Coronavirus Replication and Reverse Genetics* 287:1. Current Topics in Microbiology and Immunology.
54. **Brierley I**. (1995) Ribosomal Frameshifting on Viral RNAs. *Journal of General Virology* 76:1885.
55. **Brierley I, Boursnell ME, Binns MM, Bilimoria B, Blok VC, Brown TD, Inglis SC**. (1987) An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. *EMBO J.* 6(12):3779.
56. **Brierley I, Dos Ramos FJ**. (2006) Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Research* 119(1):29.
57. **Brierley I, Pennell S, Gilbert RJC**. (2007) Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nature Reviews Microbiology* 5(8):598.
58. **Britton P, Cavanagh D**. (2008) Nidovirus Genome Organization and Expression Mechanisms. Perlman,S., Gallagher,T., Snijder,E.J. Nidoviruses. 29. *ASM Press*.
59. **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. *Journal of Virology* 77(16):8973.
60. **Bruenn JA**. (2003) A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Research* 31(7):1821.
61. **Bull JJ, Sanjuan R, Wilke CO**. (2007) Theory of lethal mutagenesis for viruses. *Journal of Virology* 81(6):2930.
62. **Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ**. (2012) Patterns of Gene Flow Define Species of Thermophilic Archaea. *Plos Biology* 10(2).
63. **Carrillo C, Tulman ER, Delhon G, Lu Z, Carreno A, Vagnozzi A, Kutish GF, Rock DL**. (2005) Comparative genomics of foot-and-mouth disease virus. *Journal of Virology* 79(10):6487.
64. **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. *Briefings in Bioinformatics* 11(4):440.
65. **Castresana J**. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17(4):540.
66. **Cavalli-Sforza LL, Edwards AWF**. (1967) Phylogenetic Analysis Models and Estimation Procedures. *American Journal of Human Genetics* 19(3P1):233.
67. **Chan YF, Sam IC, Abubakar S**. (2010) Phylogenetic designation of enterovirus 71 genotypes and subgenotypes using complete genome sequences. *Infection Genetics and Evolution* 10(3):404.
68. **Chang HW, de Groot RJ, Egberink HF, Rottier PJM**. (2010) Feline infectious peritonitis: insights into feline coronavirus pathobiogenesis and epidemiology based on genetic analysis of the viral 3c gene. *Journal of General Virology* 91:415.
69. **Charini WA, Todd S, Gutman GA, Semler BL**. (1994) Transduction of A Human Rna Sequence by Poliovirus. *Journal of Virology* 68(10):6547.
70. **Chen HH, Kong WP, Roos RP**. (1995) The Leader Peptide of Theilers Murine Encephalomyelitis Virus Is A Zinc-Binding Protein. *Journal of Virology* 69(12):8076.

71. **Chen HH, Kong WP, Zhang L, Ward PL, Roos RP**. (1995) A Picornaviral Protein Synthesized Out of Frame with the Polyprotein Plays A Key Role in A Virus-Induced Immune-Mediated Demyelinating Disease. *Nature Medicine* 1(9):927.
72. **Chen P, Jiang M, Hu T, Liu QZ, Chen XSJ, Guo D**. (2007) Biochemical characterization of exoribonuclease encoded by SARS coronavirus. *Journal of Biochemistry and Molecular Biology* 40(5):649.
73. **Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T, Guo DY**. (2009) Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. *Proceedings of the National Academy of Sciences of the United States of America* 106(9):3484.
74. **Chen Y, Su CY, Ke M, Jin X, Xu LR, Zhang Z, Wu AD, Sun Y, Yang ZN, Tien P et al.** (2011) Biochemical and Structural Insights into the Mechanisms of SARS Coronavirus RNA Ribose 2 '-O-Methylation by nsp16/nsp10 Protein Complex. *Plos Pathogens* 7(10):e1002294.
75. **Cheng A, Zhang W, Xie Y, Jiang W, Arnold E, Sarafianos SG, Ding J**. (2005) Expression, purification, and characterization of SARS coronavirus RNA polymerase. *Virology* 335(2):165.
76. **Chirico N, Vianelli A, Belshaw R**. (2010) Why genes overlap in viruses. *Proceedings of the Royal Society B-Biological Sciences* 277(1701):3809.
77. **Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, Runckel C, Louie JK, Glaser CA, Yagi S, Schnurr DP et al.** (2008) Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections. *Proceedings of the National Academy of Sciences of the United States of America* 105(37):14124.
78. **Chow M, Newman JFE, Filman D, Hogle JM, Rowlands DJ, Brown F**. (1987) Myristylation of Picornavirus Capsid Protein Vp4 and Its Structural Significance. *Nature* 327(6122):482.
79. **Cohen JI, Rosenblum B, Ticehurst JR, Daemer RJ, Feinstone SM, Purcell RH**. (1987) Complete Nucleotide-Sequence of An Attenuated Hepatitis-A Virus - Comparison with Wild-Type Virus. *Proceedings of the National Academy of Sciences of the United States of America* 84(8):2497.
80. **Constantinou A, Davies AA, West SC**. (2001) Branch migration and Holliday junction resolution catalyzed by activities from mammalian cells. *Cell* 104(2):259.
81. **Cordey S, Gerlach D, Junier T, Zdobnov EM, Kaiser L, Tapparel C**. (2008) The cis-acting replication elements define human enterovirus and rhinovirus species. *RNA* 14(8):1568.
82. **Coutard B, Gorbalenya AE, Snijder EJ, Leontovich AM, Poupon A, De Lamballerie X, Charrel R, Gould EA, Gunther S, Norder H et al.** (2008) The VIZIER project: Preparedness against pathogenic RNA viruses. *Antiviral Research* 78(1):37.
83. **Cowley JA, Dimmock CM, Spann KM, Walker PJ**. (2000) Gill-associated virus of Penaeus monodon prawns: an invertebrate virus with ORF1a and ORF1b genes related to arteri- and coronaviruses. *Journal of General Virology* 81:1473.
84. **Cowley JA, Walker PJ**. (2002) The complete genome sequence of gill-associated virus of Penaeus monodon prawns indicates a gene organization unique among nidoviruses. *Archives of Virology* **147**(10):1977.
85. **Cowley JA, Walker PJ, Flegel TW, Lightner DV, Bonami JR, Snijder EJ, de Groot RJ**. (2012) Family *Roniviridae*. King,A.M.Q., Adams,M.J., Carstens,E.B., Lefkowitz,E.J. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. 829. Amsterdam, *Elsevier Academic Press*.

86. **Coyne JA, Orr HA**. (2004) Speciation. Sunderland, Massachusetts, U.S.A., *Sinauer Associates, Inc.*
87. **Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA**. (2009) The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research* 37:D310-D314.
88. **Culley AI, Lang AS, Suttle CA**. (2006) Metagenomic analysis of coastal RNA virus communities. *Science* 312(5781):1795.
89. **Daffis S, Szretter KJ, Schriewer J, Li JQ, Youn S, Errett J, Lin TY, Schneller S, Zust R, Dong HP et al.** (2010) 2 '-O methylation of the viral mRNA cap evades host restriction by IFIT family members. *Nature* 468(7322):452.
90. **de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, Perlman S, Poon LL, Rottier PJM, Talbot PJ et al.** (2012) Family *Coronaviridae*. King,A.M.Q., Adams,M.J., Carstens,E.B., Lefkowitz,E.J. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. 806. Amsterdam, *Elsevier Academic Press*.
91. **de Groot RJ, Cowley JA, Enjuanes L, Faaberg KS, Perlman S, Rottier PJM, Snijder EJ, Ziebuhr J, Gorbalenya AE**. (2012) Order *Nidovirales*. King,A.M.Q., Adams,M.J., Carstens,E.B., Lefkowitz,E.J. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. 785. Amsterdam, *Elsevier Academic Press*.
92. **de Jong AS, Wessels E, Dijkman HBPM, Galama JMD, Melchers WJG, Willems PHGM, van Kuppeveld FJM**. (2003) Determinants for membrane association and permeabilization of the coxsackievirus 2B protein and the identification of the Golgi complex as the target organelle. *Journal of Biological Chemistry* 278(2):1012.
93. **De Palma AM, Neyts J**. (2010) Antiviral Drugs. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 461. Washington, DC, *ASM Press*.
94. **de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H**. (2004) Classification of papillomaviruses. *Virology* 324(1):17.
95. **Decroly E, Debarnot C, Ferron F, Bouvet M, Coutard B, Imbert I, Gluais L, Papageorgiou N, Sharff A, Bricogne G et al.** (2011) Crystal Structure and Functional Analysis of the SARS-Coronavirus RNA Cap 2 '-O-Methyltransferase nsp10/nsp16 Complex. *Plos Pathogens* 7(5):e1002059.
96. **Decroly E, Imbert I, Coutard B, Bouvet ML, Selisko B, Alvarez K, Gorbalenya AE, Snijder EJ, Canard B**. (2008) Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity. *Journal of Virology* 82(16):8071.
97. **Delwart EL**. (2007) Viral metagenomics. *Reviews in Medical Virology* 17(2):115.
98. **den Boon JA, Snijder EJ, Chirnside ED, Devries AAF, Horzinek MC, Spaan WJM**. (1991) Equine Arteritis Virus Is Not A Togavirus But Belongs to the Coronaviruslike Superfamily. *Journal of Virology* 65(6):2910.
99. **Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS**. (2011) Coronaviruses An RNA proofreading machine regulates replication fidelity and diversity. *Rna Biology* 8(2):270.
100. **Denison MR, Spaan WJ, van der Meer Y, Gibson CA, Sims AC, Prentice E, Lu XT**. (1999) The putative helicase of the coronavirus mouse hepatitis virus is processed from the replicase gene polyprotein and localizes in complexes that are active in viral RNA synthesis. *Journal of Virology* 73(8):6862.

101. **Dijkstra M, Roelofsen H, Vonk RJ, Jansen RC**. (2006) Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics* 6(19):5106.
102. **Dimmic MW, Rest JS, Mindell DP, Goldstein RA**. (2002) rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution* 55(1):65.
103. **Ding CY, Zhang DB**. (2007) Molecular analysis of duck hepatitis virus type 1. *Virology* 361(1):9.
104. **Dobzhansky T**. (1937) Genetics and the Origin of Species. *Columbia University Press*.
105. **Doherty M, Todd D, McFerran N, Hoey EM**. (1999) Sequence analysis of a porcine enterovirus serotype 1 isolate: relationships with other picornaviruses. *Journal of General Virology* 80:1929.
106. **Dolja VV, Kreuze JF, Valkonen JPT**. (2006) Comparative and functional genomics of closteroviruses. *Virus Research* 117(1):38.
107. **Domingo E**. (2007) Virus Evolution. Knipe,D.M. et al. Fields Virology. 5th:389. Philadelphia, *Wolters Kluwer, Lippincott Williams & Wilkins*.
108. **Domingo E, Baranowski E, Escarmis C, Sobrino F, Holland J**. (2002) Error frequencies in picornavirus RNA polymerases: evolutionary implications for virus populations. Semler,B., Wimmer,E. Molecular Biology of Picornaviruses. 285. Washington, DC, *ASM Press*.
109. **Domingo E, Escarmis C, Baranowski E, Ruiz-Jarabo CM, Carrillo E, Nunez JI, Sobrino F**. (2003) Evolution of foot-and-mouth disease virus. *Virus Research* 91(1):47.
110. **Domingo E, Martinezsalas E, Sobrino F, Delatorre JC, Portela A, Ortin J, Lopezgalindez C, Perezbrena P, Villanueva N, Najera R et al.** (1985) The Quasispecies (Extremely Heterogeneous) Nature of Viral-Rna Genome Populations - Biological Relevance - A Review. *Gene* 40(1):1.
111. **Domingo E, Perales C, Agudo R, Arias A, Escarmis C, Ferrer-Orta C, erdaguer N**. (2010) Mutation, Quasispecies, and Lethal Mutagenisis. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 197. Washington, DC, *ASM Press*.
112. **Donaldson L, El Sayed N, Koplan J, Nduati R, Toole M, Chowdhury M, de Quadros C, Mogedal S, Singhal A** (2011) Report of the Independent Monitoring Board of the Global Polio Eradication Initiative; http://www.polioeradication.org/
113. **Donnelly MLL, Gani D, Flint M, Monaghan S, Ryan MD**. (1997) The cleavage activities of aphthovirus and cardiovirus 2A proteins. *Journal of General Virology* 78:13.
114. **Doronina VA, Wu C, de Felipe P, Sachs MS, Ryan MD, Brown JD**. (2008) Site-specific release of nascent chains from ribosomes at a sense codon. *Molecular and Cellular Biology* 28(13):4227.
115. **Dougherty JD, Park N, Gustin KE, Lloyd RE**. (2010) Interference with Cellular Gene Expression. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 165. Washington, DC, *ASM Press*.
116. **Drake JW, Charlesworth B, Charlesworth D, Crow JF**. (1998) Rates of spontaneous mutation. *Genetics* 148(4):1667.
117. **Drake JW, Holland JJ**. (1999) Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences of the United States of America* 96(24):13910.
118. **Drummond AJ, Ho SYW, Phillips MJ, Rambaut A**. (2006) Relaxed phylogenetics and dating with confidence. *Plos Biology* 4(5):699.

119. **Drummond AJ, Rambaut A**. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *Bmc Evolutionary Biology* 7.
120. **Drummond AJ, Rambaut A** (2007) Tracer; http://beast.bio.ed.ac.uk/Tracer
121. **Duffy S, Shackelton LA, Holmes EC**. (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* 9(4):267.
122. **Eck RV, Dayhoff MO**. (1966) Atlas of protein sequence and structure. Silver Springs, MD, *National Biomedical Research Foundation*.
123. **Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu XT, Scherbakova S, Graham RL, Baric RS, Stockwell TB et al.** (2010) Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. *Plos Pathogens* 6(5).
124. **Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR**. (2007) High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *Journal of Virology* 81(22):12135.
125. **Eddy SR**. (1998) Profile hidden Markov models. *Bioinformatics* 14(9):755.
126. **Edgar RC**. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792.
127. **Edwards RA, Rohwer F**. (2005) Viral metagenomics. *Nature Reviews Microbiology* 3(6):504.
128. **Egloff MP, Benarroch D, Selisko B, Romette JL, Canard B**. (2002) An RNA cap (nucleoside-2 '-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *Embo Journal* 21(11):2757.
129. **Egloff MP, Malet H, Putics A, Heinonen M, Dutartre H, Frangeul A, Gruez A, Campanacci V, Cambillau C, Ziebuhr J et al.** (2006) Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains. *Journal of Virology* 80(17):8493.
130. **Ehrenfeld E, Domingo E, Roos RP**. (2010) The Picornaviruses. *ASM Press*.
131. **Eigen M**. (1971) Selforganization of Matter and Evolution of Biological Macromolecules. *Naturwissenschaften* 58(10):465-&.
132. **Eigen M**. (1993) Viral Quasi-Species. *Scientific American* 269(1):42.
133. **Eigen M**. (1996) On the nature of virus quasispecies. *Trends in Microbiology* 4(6):216.
134. **Eigen M**. (2002) Error catastrophe and antiviral strategy. *Proceedings of the National Academy of Sciences of the United States of America* 99(21):13374.
135. **Elena SF, Sanjuan R**. (2005) Adaptive value of high mutation rates of RNA viruses: Separating causes from consequences. *Journal of Virology* 79(18):11555.
136. **Enjuanes L, Almazan F, Sola I, Zuniga S**. (2006) Biochemical aspects of coronavirus replication and virus-host interaction. *Annual Review of Microbiology* 60:211.
137. **Eriksson KK, Cervantes-Barragan L, Ludewig B, Thiel V**. (2008) Mouse Hepatitis Virus Liver Pathology Is Dependent on ADP-Ribose-1 "-Phosphatase, a Viral Function Conserved in the Alpha-Like Supergroup. *Journal of Virology* 82(24):12325.
138. **Erzberger JP, Berger JM**. (2006) Evolutionary relationships and structural mechanisms of AAA plus proteins. *Annual Review of Biophysics and Biomolecular Structure* 35:93. Annual Review of Biophysics.
139. **Etherington GJ, Dicks J, Roberts IN**. (2005) Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics* 21(3):278.

140. **Faaberg KS, Balasuriya UB, Brinton MA, Gorbalenya AE, Leung FC-C, Nauwynck H, Snijder EJ, Stadejek T, Yang H, Yoo D**. (2012) Family *Arteriviridae*. King,A.M.Q., Adams,M.J., Carstens,E.B., Lefkowitz,E.J. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. 796. Amsterdam, *Elsevier Academic Press*.
141. **Falk MM, Sobrino F, Beck E**. (1992) Vpg-Gene Amplification Correlates with Infective Particle Formation in Foot-And-Mouth-Disease Virus. *Journal of Virology* 66(4):2251.
142. **Fauquet CM, Bisaro DM, Briddon RW, Brown JK, Harrison BD, Rybicki EP, Stenger DC, Stanley J**. (2003) Revision of taxonomic criteria for species demarcation in the family Geminiviridae, and an updated list of begomovirus species. *Archives of Virology* 148(2):405.
143. **Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA**. (2005) Eighth Report on the International Committee on Taxonomy of Viruses. Fauquet,C.M. et al. London, San Diego, *Elsevier Academic Press*.
144. **Feinstone SM, Kapikian AZ, Purcell RH**. (1973) Hepatitis A - Detection by Immune Electron-Microscopy of A Viruslike Antigen Associated with Acute Illness. *Science* 182(4116):1026.
145. **Felsenstein J**. (1981) Evolutionary Trees from Dna-Sequences - A Maximum-Likelihood Approach. *Journal of Molecular Evolution* 17(6):368.
146. **Felsenstein J**. (1985) Phylogenies and the Comparative Method. *American Naturalist* 125(1):1.
147. **Feng DF, Doolittle RF**. (1987) Progressive Sequence Alignment As A Prerequisite to Correct Phylogenetic Trees. *Journal of Molecular Evolution* 25(4):351.
148. **Ferrer-Orta C, Arias A, Escarmis C, Verdaguer N**. (2006) A comparison of viral RNA-dependent RNA polymerases. *Current Opinion in Structural Biology* 16(1):27.
149. **Ferrer-Orta C, Verdaguer N**. (2009) RNA virus polymerases. Cameron,C.E., Gotte,M., Raney,K.D. Viral Genome Replication. 383. New York, NY, *Springer*.
150. **Ferron F, Longhi S, Henrissat B, Canard B**. (2002) Viral RNA-polymerases - a predicted 2 '-O-ribose methyltransferase domain shared by all Mononegavirales. *Trends in Biochemical Sciences* 27(5):222.
151. **Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL et al.** (2008) The Pfam protein families database. *Nucleic Acids Research* 36:D281-D288.
152. **Fitch WM**. (1971) Toward Defining Course of Evolution - Minimum Change for A Specific Tree Topology. *Systematic Zoology* 20(4):406-&.
153. **Forterre P**. (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus Research* 117(1):5.
154. **Franssen H, Leunissen J, Goldbach R, Lomonossoff G, Zimmern D**. (1984) Homologous sequences in non-structural proteins from cowpea mosaic virus and picornaviruses. *EMBO J.* 3:855.
155. **Gago S, Elena SF, Flores R, Sanjuan R**. (2009) Extremely High Mutation Rate of a Hammerhead Viroid. *Science* 323(5919):1308.
156. **Gerber K, Wimmer E, Paul AV**. (2001) Biochemical and Genetic Studies of the Initiation of Human Rhinovirus 2 RNA Replication: Identification of a cis-Replicating Element in the Coding Sequence of 2A(pro). *Journal of Virology* 75(22):10979.
157. **Gerstein M, Sonnhammer ELL, Chothia C**. (1994) Volume Changes in Protein Evolution. *Journal of Molecular Biology* 236(4):1067.

158. **Ghosh RC, Ball BV, Willcocks MM, Carter MJ**. (1999) The nucleotide sequence of sacbrood virus of the honey bee: an insect picorna-like virus. *Journal of General Virology* 80:1541.
159. **Gilbert W**. (1986) Origin of Life - the RNA World. *Nature* 319(6055):618.
160. **Goldbach R**. (1987) Genome similarities between plant and animal RNA viruses. *Microbiol.Sci.* 4:197.
161. **Goldfarb LG, Gajdusek DC**. (1992) Viliuisk Encephalomyelitis in the Iakut People of Siberia. *Brain* 115:961.
162. **Gonzaalez JM, Gomez-Puertas P, Cavanagh D, Gorbalenya AE, Enjuanes L**. (2003) A comparative sequence analysis to revise the current taxonomy of the family Coronaviridae. *Archives of Virology* 148(11):2207.
163. **Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ**. (2000) Identification of a cis-acting replication element within the poliovirus coding region. *Journal of Virology* 74(10):4590.
164. **Goodfellow IG, Kerrigan D, Evans DJ**. (2003) Structure and function analysis of the poliovirus cis-acting replication element (CRE). *Rna-A Publication of the Rna Society* 9(1):124.
165. **Goodman SN**. (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* 130(12):1005.
166. **Gorbalenya AE**. (1992) Host-related sequences in RNA virus genomes. *Semin.Virol.* 3:359.
167. **Gorbalenya AE**. (1995) Origin of RNA viral genomes; Approaching the problem by comparative sequence analysis. Gibbs,A.J., Calisher,C.H., Garcia-Arenal,F. Molecular Basis of Virus Evolution. 49. Cambridge, UK, *Cambridge University Press*.
168. **Gorbalenya AE**. (2000) Papain-like fold, acyl-enzyme intermediate and a complex evolution history are predicted for 2A proteins of several picornaviruses from bioinformatics analysis of their distant relationships. *Abstracts of XIth Meeting of the European Study Group on Molecular Biology of Picornaviruses* :J13. Baia delle Zagare. 2000)
169. **Gorbalenya AE**. (2001) Big nidovirus genome - When count and order of domains matter. *Nidoviruses (Coronaviruses and Arteriviruses)* 494:1. Advances in Experimental Medicine and Biology.
170. **Gorbalenya AE**. (2008) Genomics and Evolution of the Nidovirales. Perlman,S., Gallagher,T., Snijder,E.J. Nidoviruses. 15. Washington, DC, *ASM Press*.
171. **Gorbalenya AE, Chumakov KM, Agol VI**. (1978) RNA-binding properties of nonstructural polypeptide G of encephalomyocarditis virus. *Virology* 88(1):183.
172. **Gorbalenya AE, Donchenko AP, Blinov VM**. (1986) A possible common origin of poliovirus proteins with different functions. *Molekularnaya Genetika, Microbiologiya, i Virusologiya* (1):36.
173. **Gorbalenya AE, Donchenko AP, Blinov VM, Koonin EV**. (1989) Cysteine Proteases of Positive Strand Rna Viruses and Chymotrypsin-Like Serine Proteases - A Distinct Protein Superfamily with A Common Structural Fold. *Febs Letters* 243(2):103.
174. **Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ**. (2006) Nidovirales: Evolving the largest RNA virus genome. *Virus Research* 117(1):17.
175. **Gorbalenya AE, Koonin E**. (1993) Comparative analysis of amino-acid sequences of key enzymes of replication and expression of positive-strand RNA viruses: validity of approach and functional and evolutionary implications. *Sov.Sci.Rev.D Physicochem.Biol.* 11:1.

176. **Gorbalenya AE, Koonin EV**. (1989) Viral-Proteins Containing the Purine Ntp-Binding Sequence Pattern. *Nucleic Acids Research* 17(21):8413.
177. **Gorbalenya AE, Koonin EV**. (1993) Helicases - Amino-Acid-Sequence Comparisons and Structure-Function-Relationships. *Current Opinion in Structural Biology* 3(3):419.
178. **Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM**. (1988) A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEBS Lett.* 235:16.
179. **Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM**. (1989) Coronavirus Genome - Prediction of Putative Functional Domains in the Non-Structural Polyprotein by Comparative Amino-Acid Sequence-Analysis. *Nucleic Acids Research* 17(12):4847.
180. **Gorbalenya AE, Koonin EV, Lai MMC**. (1991) Putative papain-related thiol proteases of positive-strand RNA viruses. *FEBS Lett.* 288:201.
181. **Gorbalenya AE, Koonin EV, Wolf YA**. (1990) A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS Lett.* **262**:145.
182. **Gorbalenya AE, Lauber C**. (2010) Origin and Evolution of the Picornaviridae Proteome. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 253. Washington, *ASM Press*.
183. **Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM et al.** (2010) Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antiviral Research* 87(2):95.
184. **Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KHJ, Ward VK**. (2002) The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. *Journal of Molecular Biology* 324(1):47.
185. **Gorbalenya AE, Snijder EJ**. (1996) Viral cysteine proteinases. *Perspectives in Drug Discovery and Design* 6:64.
186. **Gorbalenya AE, Svitkin YV, Kazachkov YA, Agol VI**. (1979) Encephalomyocarditis virus-specific polypeptide p22 is involved in the processing of the viral precursor polypeptides. *FEBS Lett.* 108(1):1.
187. **Graham RL, Sparks JS, Eckerle LD, Sims AC, Denison MR**. (2008) SARS coronavirus replicase proteins in pathogenesis. *Virus Research* 133(1):88.
188. **Grant BJ, Rodrigues APC, Elsawy KM, McCammon JA, Caves LSD**. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22(21):2695.
189. **Green KY, Ando T, Balayan MS, Berke T, Clarke IN, Estes MK, Matson DO, Nakata S, Neill JD, Studdert MJ et al.** (2000) Taxonomy of the caliciviruses. *Journal of Infectious Diseases* 181:S322-S330.
190. **Greninger AL, Runckel C, Chiu CY, Haggerty T, Parsonnet J, Ganem D, Derisi JL**. (2009) The complete genome of klassevirus - a novel picornavirus in pediatric stool. *Virology Journal* 6.
191. **Gribskov M, Mclachlan AD, Eisenberg D**. (1987) Profile Analysis - Detection of Distantly Related Proteins. *Proceedings of the National Academy of Sciences of the United States of America* 84(13):4355.

192. **Gromeier M, Wimmer E, Gorbalenya AE**. (1999) Genetics, Pathogenesis and Evolution of picornaviruses. Domingo,E., Webster,R.G., Holland,J.J. Origin and Evolution of Viruses. (12):287. San Diego, *Academic Press*.
193. **Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ et al.** (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. *Science* 302(5643):276.
194. **Guarino LA, Bhardwaj K, Dong W, Sun JC, Holzenburg A, Kao C**. (2005) Mutational analysis of the SARS virus Nsp15 endoribonuclease: Identification of residues affecting hexamer formation. *Journal of Molecular Biology* 353(5):1106.
195. **Guarne A, Tormo J, Kirchweger R, Pfistermueller D, Fita I, Skern T**. (1998) Structure of the foot-and-mouth disease virus leader protease: a papain- like fold adapted for self-processing and eIF4G recognition. *EMBO J.* 17(24):7469.
196. **Guindon S, Gascuel O**. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5):696.
197. **Hales LM, Knowles NJ, Reddy PS, Xu L, Hay C, Hallenbeck PL**. (2008) Complete genome sequence analysis of Seneca Valley virus-001, a novel oncolytic picornavirus. *Journal of General Virology* 89:1265.
198. **Hanada K, Suzuki Y, Gojobori T**. (2004) A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular Biology and Evolution* 21(6):1074.
199. **Hansen JL, Long AM, Schultz SC**. (1997) Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure* 5(8):1109.
200. **Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, Smith CM, Rota PA, Baker SC**. (2004) Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. *Journal of Virology* 78(24):13600.
201. **Hartigan JA**. (1973) Minimum evolution fits to a given tree. *Biometrics* 29:53.
202. **Hastie KM, Kimberlin CR, Zandonatti MA, Macrae IJ, Saphire EO**. (2011) Structure of the Lassa virus nucleoprotein reveals a dsRNA-specific 3 ' to 5 ' exonuclease activity essential for immune suppression. *Proceedings of the National Academy of Sciences of the United States of America* 108(6):2396.
203. **Hazelton PR, Gelderblom HR**. (2003) Electron microscopy for rapid diagnosis of infectious agents in emergent situations. *Emerging Infectious Diseases* 9(3):294.
204. **Hebert PDN, Ratnasingham S, deWaard JR**. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences* 270:S96-S99.
205. **Hellen CUT, de Breyne S**. (2007) A distinct group of hepacivirus/pestivirus-like internal ribosomal entry sites in members of diverse Picornavirus genera: Evidence for modular exchange of functional noncoding RNA elements by recombination. *Journal of Virology* 81(11):5850.
206. **Hemelt IE, Huxsoll DL, Warner AR**. (1974) Comparison of Mhg Virus with Mouse Encephalomyelitis Viruses. *Laboratory Animal Science* 24(3):523.
207. **Hendry AP**. (2009) Evolutionary Biology Speciation. *Nature* 458(7235):162.
208. **Hendry E, Hatanaka H, Fry E, Smyth M, Tate J, Stanway G, Santti J, Maaronen M, Hyypia T, Stuart D**. (1999) The crystal structure of coxsackievirus A9: new insights into the uncoating mechanisms of enteroviruses. *Structure* 7(12):1527.
209. **Henikoff S, Henikoff JG**. (1994) Position-Based Sequence Weights. *Journal of Molecular Biology* 243(4):574.

210. **Herold J, Siddell SG, Gorbalenya AE**. (1999) A human RNA viral cysteine proteinase that depends upon a unique Zn2+-binding finger connecting the two domains of a papain-like fold. *Journal of Biological Chemistry* 274(21):14918.
211. **Hicks AL, Duffy S**. (2011) Genus-Specific Substitution Rate Variability among Picornaviruses. *Journal of Virology* 85(15):7942.
212. **Hodgman TC**. (1988) A new superfamily of replicative proteins. *Nature* 333(333):22.
213. **Hofmann K, Stoffel W**. (1993) TMbase - A database of membrane spanning protein segments. *Biol Chem Hoppe-Seyler* 373:166.
214. **Hogle JM, Chow M, Filman DJ**. (1985) 3-Dimensional Structure of Poliovirus at 2.9 A Resolution. *Science* 229(4720):1358.
215. **Hollister JR, Vagnozzi A, Knowles NJ, Rieder E**. (2008) Molecular and phylogenetic analyses of bovine rhinovirus type 2 shows it is closely related to foot-and-mouth disease virus. *Virology* 373(2):411.
216. **Holmes EC**. (2003) Error thresholds and the constraints to RNA virus evolution. *Trends in Microbiology* 11(12):543.
217. **Holmes EC**. (2009) The Evolution and Emergence of RNA viruses. New York, *Oxford University Press*.
218. **Holmes EC**. (2009) The Evolutionary Genetics of Emerging Viruses. *Annual Review of Ecology Evolution and Systematics* 40:353. Annual Review of Ecology Evolution and Systematics.
219. **Holmes EC**. (2010) The RNA Virus Quasispecies: Fact or Fiction? *Journal of Molecular Biology* 400(3):271.
220. **Holmes EC**. (2011) What Does Virus Evolution Tell Us about Virus Origins? *Journal of Virology* 85(11):5247.
221. **Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D**. (2008) Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. *Virology Journal* 5.
222. **Holtz LR, Finkbeiner SR, Zhao GY, Kirkwood CD, Girones R, Pipas JM, Wang D**. (2009) Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. *Virology Journal* 6.
223. **Hraber P, Kuiken C, Waugh M, Geer S, Bruno WJ, Leitner T**. (2008) Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index. *Journal of General Virology* 89:2098.
224. **Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S**. (2011) Alphacoronavirus Transmissible Gastroenteritis Virus nsp1 Protein Suppresses Protein Translation in Mammalian Cells and in Cell-Free HeLa Cell Extracts but Not in Rabbit Reticulocyte Lysate. *Journal of Virology* 85(1):638.
225. **Hughes AL**. (2004) Phylogeny of the Picornaviridae and differential evolutionary divergence of picornavirus proteins. *Infect.Genet.Evol.* 4:143.
226. **Hughes PJ, Stanway G**. (2000) The 2A proteins of three diverse picornaviruses are related to each other and to the H-rev107 family of proteins involved in the control of cell proliferation. *Journal of General Virology* 81(Pt 1):201.
227. **Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ**. (2008) The 20 years of PROSITE. *Nucleic Acids Res.* 36(Database issue):D245-D249.
228. **Igarashi A**. (1978) Isolation of A Singhs Aedes-Albopictus Cell Clone Sensitive to Dengue and Chikungunya Viruses. *Journal of General Virology* 40(SEP):531.

229. **Imbert I, Guillemot JC, Bourhis JM, Bussetta C, Coutard B, Egloff MP, Ferron F, Gorbalenya AE, Canard B**. (2006) A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus. *Embo Journal* 25(20):4933.
230. **Imbert I, Snijder EJ, Dimitrova M, Guillemot JC, Lecine P, Canard B**. (2008) The SARS-coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. *Virus Research* 133(2):136.
231. **Isawa H, Asano S, Sahara K, Iizuka T, Bando H**. (1998) Analysis of genetic information of an insect picorna-like virus, infectious flacherie virus of silkworm: evidence for evolutionary relationships among insect, mammalian and plant picorna(-like) viruses. *Archives of Virology* 143(1):127.
232. **Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, Ziebuhr J**. (2004) Major genetic marker of nidoviruses encodes a replicative endoribonuclease. *Proceedings of the National Academy of Sciences of the United States of America* 101(34):12694.
233. **Ivanovsky D**. (1892) Concerning the mosaic disease of the tobacco plant. *St.Petersb.Acad.Imp.Sci.Bul.* 35:67.
234. **Jackson DJ, Gangnon RE, Evans MD, Roberg KA, Anderson EL, Pappas TE, Printz MC, Lee WM, Shult PA, Reisdorf E et al.** (2008) Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children. *American Journal of Respiratory and Critical Care Medicine* 178(7):667.
235. **Jegouic S, Joffret ML, Blanchard C, Riquet FB, Perret C, Pelletier I, Colbere-Garapin F, Rakoto-Andrianarivelo M, Delpeyroux F**. (2009) Recombination between Polioviruses and Co-Circulating Coxsackie A Viruses: Role in the Emergence of Pathogenic Vaccine-Derived Polioviruses. *Plos Pathogens* 5(5).
236. **Jenkins GM, Rambaut A, Pybus OG, Holmes EC**. (2002) Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *Journal of Molecular Evolution* 54(2):156.
237. **Jenkins GM, Worobey M, Woelk CH, Holmes EC**. (2001) Nonquasispecies evidence for the evolution of RNA viruses. *Molecular Biology and Evolution* 18(6):987.
238. **Jiang P, Faase JAJ, Toyoda H, Paul A, Wimmer E, Gorbalenya AE**. (2007) Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. *Proceedings of the National Academy of Sciences of the United States of America* 104(22):9457.
239. **Johansson S, Niklasson B, Maizel J, Gorbalenya AE, Lindberg AM**. (2002) Molecular analysis of three Ljungan virus isolates reveals a new, close-to-root lineage of the Picornaviridae with a cluster of two unrelated 2A proteins. *Journal of Virology* 76(17):8920.
240. **Jones DT**. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292(2):195.
241. **Jones MS, Lukashov VV, Ganac RD, Schnurr DP**. (2007) Discovery of a novel human picornavirus in a stool sample from a pediatric patient presenting with fever of unknown origin. *Journal of Clinical Microbiology* 45(7):2144.
242. **Joseph JS, Saikatendu KS, Subramanian V, Neuman BW, Buchmeier MJ, Stevens RC, Kuhn P**. (2007) Crystal structure of a monomeric form of severe acute respiratory syndrome coronavirus endonuclease nsp15 suggests a role for hexamerization as an allosteric switch. *Journal of Virology* 81(12):6700.
243. **Junglen S, Kurth A, Kuehl H, Quan PL, Ellerbrok H, Pauli G, Nitsche A, Nunn C, Rich SM, Lipkin WI et al.** (2009) Examining Landscape Factors Influencing

Relative Distribution of Mosquito Genera and Frequency of Virus Infection. *Ecohealth* 6(2):239.

244. **Kadare G, Haenni AL**. (1997) Virus-encoded RNA helicases. *Journal of Virology* 71(4):2583.
245. **Kamer G, Argos P**. (1984) Primary Structural Comparison of RNA-Dependent Polymerases from Plant, Animal and Bacterial-Viruses. *Nucl.Acids Res.* 12(18):7269.
246. **Kamitani W, Narayanan K, Huang C, Lokugamage K, Ikegami T, Ito N, Kubo H, Makino S**. (2006) Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proceedings of the National Academy of Sciences of the United States of America* 103(34):12885.
247. **Kampstra P**. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets* 28:1.
248. **Kang H, Bhardwaj K, Li Y, Palaninathan S, Sacchettini J, Guarino L, Leibowitz JL, Kao CC**. (2007) Biochemical and genetic analyses of murine hepatitis virus nsp15 endoribonuclease. *Journal of Virology* 81(24):13587.
249. **Kapoor A, Victoria J, Simmonds P, Slikas E, Chieochansin T, Naeem A, Shaukat S, Sharif S, Alam MM, Angez M et al.** (2008) A highly prevalent and genetically diversified Picornaviridae genus in South Asian children. *Proceedings of the National Academy of Sciences of the United States of America* 105(51):20482.
250. **Kapoor A, Victoria J, Simmonds P, Wang C, Shafer RW, Nims R, Nielsen O, Delwart E**. (2008) A highly divergent picornavirus in a marine mammal. *Journal of Virology* 82(1):311.
251. **Keeney S, Giroux CN, Kleckner N**. (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88(3):375.
252. **Keese PK, Gibbs A**. (1992) Origins of Genes - Big-Bang Or Continuous Creation. *Proceedings of the National Academy of Sciences of the United States of America* 89(20):9489.
253. **Kim KH, Lommel SA**. (1994) Identification and Analysis of the Site of -1 Ribosomal Frameshifting in Red-Clover Necrotic Mosaic-Virus. *Virology* 200(2):574.
254. **Kim KH, Lommel SA**. (1998) Sequence element required for efficient -1 ribosomal frameshifting in red clover necrotic mosaic dianthovirus. *Virology* 250(1):50.
255. **Kim MC, Kwon YK, Joh SJ, Lindberg AM, Kwon JH, Kim JH, Kim SJ**. (2006) Molecular analysis of duck hepatitis virus type 1 reveals a novel lineage close to the genus Parechovirus in the family Picornaviridae. *Journal of General Virology* 87:3307.
256. **Kimura M**. (1983) The neutral theory of molecular evolution. *Cambridge University Press*.
257. **King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ**. (2012) Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. *Academic Press*.
258. **King AMQ, Brown QF, Christian P, Hovi T, Hypia T, Knowles NJ, Lemon SM, Minor PD, Palmenberg AC, Skern T et al.** (2000) Family *Picornaviridae*. Van Regenmortel,M.H.V. et al. Virus Taxonomy, Seventh Report of the Internation Committee on Taxonomy of Viruses. 657. New York, NY, *Academic Press*.
259. **Kingsbury DW**. (1985) Species Classification Problems in Virus Taxonomy. *Intervirology* 24(2):62.

260. **Kirkegaard K, Baltimore D**. (1986) The Mechanism of Rna Recombination in Poliovirus. *Cell* 47(3):433.
261. **Kirkegaard K, Semler B**. (2010) Genome Replication II: the Process. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 127. Washington, DC, *ASM Press*.
262. **Kitchen A, Shackelton LA, Holmes EC**. (2011) Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proceedings of the National Academy of Sciences of the United States of America* 108(1):238.
263. **Knowles NJ, Hovi T, Hyypia T, King AMQ, Lindberg AM, Pallansch MA, Palmenberg AC, Simmonds P, Skern T, Stanway G et al.** (2012) Family *Picornaviridae*. King,A.M.Q., Adams,M.J., Carstens,E.B., Lefkowitz,E.J. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. 855. *Academic Press*.
264. **Knowles NJ, Hovi T, King AMQ, Stanway G**. (2010) Overview of Taxonomy. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 19. Washington, *ASM Press*.
265. **Knowles NJ, Samuel AR**. (2003) Molecular epidemiology of foot-and-mouth disease virus. *Virus Research* 91(1):65.
266. **Kong WP, Ghadge GD, Roos RP**. (1994) Involvement of Cardiovirus Leader in Host Cell-Restricted Virus Expression. *Proceedings of the National Academy of Sciences of the United States of America* 91(5):1796.
267. **Kong WP, Roos RP**. (1991) Alternative Translation Initiation Site in the DA Strain of Theilers Murine Encephalomyelitis Virus. *Journal of Virology* 65(6):3395.
268. **Koonin EV**. (1991) The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *Journal of General Virology* **72**:2197.
269. **Koonin EV, Gorbalenya AE**. (1992) An insect picornavirus may have genome organization similar to that of caliciviruses. *FEBS Lett.* 297(1-2):81.
270. **Koonin EV, Wolf YI, Nagasaki K, Dolja VV**. (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nature Rev.Microb.* 6(12):925.
271. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW**. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution* 23(10):1891.
272. **Krakauer DC**. (2000) Stability and evolution of overlapping genes. *Evolution* 54(3):731.
273. **Krogh A, Brown M, Mian IS, Sjolander K, Haussler D**. (1994) Hidden Markov-Models in Computational Biology - Applications to Protein Modeling. *Journal of Molecular Biology* 235(5):1501.
274. **Krumbholz A, Dauber M, Henke A, Birch-Hirschfeld E, Knowles NJ, Stelzner A, Zell R**. (2002) Sequencing of porcine enterovirus groups II and III reveals unique features of both virus groups. *Journal of Virology* 76(11):5813.
275. **Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K**. (2011) Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution* 29(2):457.
276. **Kun A, Santos M, Szathmary E**. (2005) Real ribozymes suggest a relaxed error threshold. *Nat.Genet.* 37(9):1008.
277. **Laine P, Savolainen C, Blomqvist S, Hovi T**. (2005) Phylogenetic analysis of human rhinovirus capsid protein VP1 and 2A protease coding sequences confirms shared genus-like relationships with human enteroviruses. *Journal of General Virology* 86:697.
278. **Lander ES**. (2011) Initial impact of the sequencing of the human genome. *Nature* 470(7333):187.

279. **Lang AS, Culley AI, Suttle CA**. (2004) Genome sequence and characterization of a virus (HaRNAV) related to picoma-like viruses that infects the marine toxic bloom-forming alga Heterosigma akashiwo. *Virology* 320(2):206.
280. **Lanzi G, de Miranda JR, Boniotti MB, Cameron CE, Lavazza A, Capucci L, Camazine SM, Rossi C**. (2006) Molecular and biological characterization of deformed wing virus of honeybees (Apis mellifera L.). *Journal of Virology* 80(10):4998.
281. **Lauber C, Gorbalenya AE**. (2012) Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses* 4:1425.
282. **Lauber C, Gorbalenya AE**. (2012) Partitioning the Genetic Diversity of a Virus Family: Approach and Evaluation through a Case Study of Picornaviruses. *Journal of Virology* 86(7):3890.
283. **Lauber C, Gorbalenya AE**. (2012) Toward Genetics-Based Virus Taxonomy: Comparative Analysis of a Genetics-Based Classification and the Taxonomy of Picornaviruses. *Journal of Virology* 86(7):3905.
284. **Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE**. (2012) Mesoniviridae: a proposed new family in the order Nidovirales formed by a single species of mosquito-borne viruses. *Archives of Virology* :DOI: 10.1007/s00705-012-1295-x.
285. **Lauring AS, Andino R**. (2010) Quasispecies Theory and the Behavior of RNA Viruses. *Plos Pathogens* 6(7).
286. **Lavergne P, Vuong QH**. (1996) Nonparametric selection of regressors: The nonnested case. *Econometrica* 64(1):207.
287. **Le Gall O, Christian P, Fauquet CM, King AMQ, Knowles NJ, Nakashima N, Stanway G, Gorbalenya AE**. (2008) Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T=3 virion architecture. *Archives of Virology* 153(4):715.
288. **Ledford RM, Patel NR, Demenczuk TM, Watanyar A, Herbertz T, Collett MS, Pevear DC**. (2004) VP1 sequencing of all human rhinovirus serotypes: Insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds. *Journal of Virology* 78(7):3663.
289. **Lefebure T, Douady CJ, Gouy M, Gibert J**. (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: Proposal of a molecular threshold to help species delimitation. *Molecular Phylogenetics and Evolution* 40(2):435.
290. **Lefkowitz EJ, Wang C, Upton C**. (2006) Poxviruses: past, present and future. *Virus Research* 117(1):105.
291. **Levy H, Bostina M, Filman DJ, Hogle JM**. (2010) Cell Entry: a Biochemical and Structural Perspective. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 87. Washington, DC, *ASM Press*.
292. **Lewis-Rogers N, Bendall ML, Crandall KA**. (2009) Phylogenetic Relationships and Molecular Adaptation Dynamics of Human Rhinoviruses. *Molecular Biology and Evolution* 26(5):969.
293. **Lewis-Rogers N, Crandall KA**. (2010) Evolution of Picornaviridae: An examination of phylogenetic relationships and cophylogeny. *Molecular Phylogenetics and Evolution* 54(3):995.
294. **Lewis-Rogers N, McClellan DA, Crandall KA**. (2008) The evolution of foot-and-mouth disease virus: Impacts of recombination and selection. *Infection Genetics and Evolution* 8(6):786.

295. **Li F, Browning GF, Studdert MJ, Crabb BS**. (1996) Equine rhinovirus 1 is more closely related to foot-and-mouth disease virus than to other picornaviruses. *Proceedings of the National Academy of Sciences of the United States of America* 93(3):990.
296. **Li JP, Baltimore D**. (1988) Isolation of Poliovirus 2C Mutants Defective in Viral-Rna Synthesis. *Journal of Virology* 62(11):4016.
297. **Li JP, Baltimore D**. (1990) An Intragenic Revertant of A Poliovirus-2C Mutant Has An Uncoating Defect. *Journal of Virology* 64(3):1102.
298. **Li LL, Victoria J, Kapoor A, Blinkova O, Wang CL, Babrzadeh F, Mason CJ, Pandey P, Triki H, Bahri O et al.** (2009) A Novel Picornavirus Associated with Gastroenteritis. *Journal of Virology* 83(22):12002.
299. **Liang Z, Kumar ASM, Jones MS, Knowles NJ, Lipton HL**. (2008) Phylogenetic Analysis of the Species Theilovirus: Emerging Murine and Human Pathogens. *Journal of Virology* 82(23):11545.
300. **Liljas L, Tate J, Lin T, Christian P, Johnson JE**. (2002) Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses and insect picorna-like viruses. *Archives of Virology* 147(1):59.
301. **Loeffler F, Frosch P**. (1898) Bericht der Kommission zur Erforschung der Maul- und Klauenseuche bei dem Institut für Infektionskrankheiten in Berlin. *Zentralbl.f.Bakteriol.I.Abt.* 23:371.
302. **Lorusso A, Decaro N, Schellen P, Rottier PJM, Buonavoglia C, Haijema BJ, de Groot RJ**. (2008) Gain, Preservation, and Loss of a Group 1a Coronavirus Accessory Glycoprotein. *Journal of Virology* 82(20):10312.
303. **Lu HH, Li XY, Cuconati A, Wimmer E**. (1995) Analysis of Picornavirus 2A(Pro) Proteins - Separation of Proteinase from Translation and Replication Functions. *Journal of Virology* 69(12):7445.
304. **Lu HH, Wimmer E**. (1996) Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. *Proceedings of the National Academy of Sciences of the United States of America* 93(4):1412.
305. **Lukashev AN**. (2010) Recombination among picornaviruses. *Reviews in Medical Virology* 20(5):327.
306. **Luke GA, de Felipe P, Lukashev A, Kallioinen SE, Bruno EA, Ryan MD**. (2008) Occurrence, function and evolutionary origins of '2A-like' sequences in virus genomes. *Journal of General Virology* 89:1036.
307. **Lynch M**. (2010) Evolution of the mutation rate. *Trends Genet.* 26(8):345.
308. **Lynch M, Conery JS**. (2003) The origins of genome complexity. *Science* 302(5649):1401.
309. **Maes P, Klempa B, Clement J, Matthijnssens J, Gajdusek DC, Kruger DH, Van Ranst M**. (2009) A proposal for new criteria for the classification of hantaviruses, based on S and M segment protein sequences. *Infection Genetics and Evolution* 9(5):813.
310. **Marc D, Drugeon G, Haenni AL, Girard M, Vanderwerf S**. (1989) Role of Myristoylation of Poliovirus Capsid Protein Vp4 As Determined by Site-Directed Mutagenesis of Its N-Terminal Sequence. *EMBO J.* 8(9):2661.
311. **Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P**. (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19):2462.
312. **Martin-Belmonte F, Lopez-Guerrero JA, Carrasco L, Alonso MA**. (2000) The amino-terminal nine amino acid sequence of poliovirus capsid VP4 protein is

sufficient to confer N-myristoylation and targeting to detergent-insoluble membranes. *Biochemistry* 39(5):1083.

313. **Martinez-Salas E, Ryan MD**. (2010) Translation and protein processing. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 141. Washington, *ASM Press*.
314. **Marvil P, Knowles NJ, Mockett APA, Britton P, Brown TDK, Cavanagh D**. (1999) Avian encephalomyelitis virus is a picornavirus and is most closely related to hepatitis A virus. *Journal of General Virology* 80:653.
315. **Mason PW, Rieder E, Baxt B**. (1994) RGD Sequence of Foot-And-Mouth-Disease Virus Is Essential for Infecting Cells Via the Natural Receptor But Can be Bypassed by An Antibody-Dependent Enhancement Pathway. *Proceedings of the National Academy of Sciences of the United States of America* 91(5):1932.
316. **Masters PS**. (2006) The molecular biology of coronaviruses. *Advances in Virus Research, Vol 66* 66:193. Advances in Virus Research.
317. **Matthijnssens J, Ciarlet M, Heiman E, Arijs I, Delbeke T, McDonald SM, Palombo EA, Iturriza-Gomara M, Maes P, Patton JT 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. *Journal of Virology* 82(7):3204.
318. **McIntyre CL, Leitch ECM, Savolainen-Kopra C, Hovi T, Simmonds P**. (2010) Analysis of Genetic Diversity and Sites of Recombination in Human Rhinovirus Species C. *Journal of Virology* 84(19):10297.
319. **McKnight KL, Lemon SM**. (1998) The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *Rna-A Publication of the Rna Society* 4(12):1569.
320. **Meyer CP, Paulay G**. (2005) DNA barcoding: Error rates based on comprehensive sampling. *Plos Biology* 3(12):2229.
321. **Mi S, Durbin R, Huang HV, Rice CM, Stollar V**. (1989) Association of the Sindbis Virus-RNA Methyltransferase Activity with the Nonstructural Protein-Nsp1. *Virology* 170(2):385.
322. **Milne RG**. (1984) The Species Problem in Plant Virology. *Microbiological Sciences* 1(5):113.
323. **Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J**. (2006) Discovery of an RNA virus 3 '-> 5 ' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America* 103(13):5108.
324. **Mitra T, Sosnovtsev SV, Green KY**. (2004) Mutagenesis of Tyrosine 24 in the VPg Protein Is Lethal for Feline Calicivirus. *Journal of Virology* 78(9):4931.
325. **Morace G, Kusov Y, Dzagurov G, Beneduce F, Gauss-Muller V**. (2008) The unique role of domain 2A of the hepatitis A virus precursor polypeptide P1-2A in viral morphogenesis. *BMB.Rep.* 41(9):678.
326. **Moratorio G, Costa-Mattioli M, Piovani R, Romero H, Musto H, Cristina J**. (2007) Bayesian coalescent inference of hepatitis A virus populations: evolutionary rates and patterns. *Journal of General Virology* 88:3039.
327. **Morran LT, Schmidt OG, Gelarden IA, Parrish RC, Lively CM**. (2011) Running with the Red Queen: Host-Parasite Coevolution Selects for Biparental Sex. *Science* 333(6039):216.
328. **Murzin AG, Brenner SE, Hubbard T, Chothia C**. (1995) SCOP - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* 247(4):536.

329. **Nagy PD, Simon AE**. (1997) New insights into the mechanisms of RNA recombination. *Virology* 235(1):1.
330. **Nakashima N, Shibuya N**. (2006) Multiple coding sequences for the genome-linked virus protein (VPg) in dicistroviruses. *Journal of Invertebrate Pathology* 92(2):100.
331. **Nature Medicine Editorial**. (2012) The persistence of polio. *Nature Medicine* 18:323.
332. **Nedialkova DD, Gorbalenya AE, Snijder EJ**. (2010) Arterivirus Nsp1 Modulates the Accumulation of Minus-Strand Templates to Control the Relative Abundance of Viral mRNAs. *Plos Pathogens* 6(2).
333. **Nedialkova DD, Ulferts R, van den Born E, Lauber C, Gorbalenya AE, Ziebuhr J, Snijder EJ**. (2009) Biochemical Characterization of Arterivirus Nonstructural Protein 11 Reveals the Nidovirus-Wide Conservation of a Replicative Endoribonuclease. *Journal of Virology* 83(11):5671.
334. **Neuman BW, Joseph JS, Saikatendu KS, Serrano P, Chatterjee A, Johnson MA, Liao L, Klaus JP, Yates JR, Wuethrich K et al.** (2008) Proteornics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *Journal of Virology* 82(11):5279.
335. **Nga PT, KieuAnh NT, Cuong VD, Nam VS, Hang PTM, et al.** (2002) Surveillance of Japanse encephalitis in Vietnam (2000-2001). *J Hyg Epidemiol* XII:5.
336. **Nga PT, Parquet MD, Lauber C, Parida M, Nabeshima T, Yu FX, Thuy NT, Inoue S, Ito T, Okamoto K et al.** (2011) Discovery of the First Insect Nidovirus, a Missing Evolutionary Link in the Emergence of the Largest RNA Virus Genomes. *Plos Pathogens* 7(9):e1002215.
337. **Nurmemmedov E, Castelnovo M, Catalano CE, Evilevitch A**. (2007) Biophysics of viral infectivity: matching genome length with capsid size. *Q.Rev.Biophys.* 40(4):327.
338. **Oberste MS, Jiang X, Maher K, Nix WA, Jiang BM**. (2008) The complete genome sequences for three simian enteroviruses isolated from captive primates. *Archives of Virology* 153(11):2117.
339. **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. *Journal of Virology* 73(3):1941.
340. **Oberste MS, Maher K, Pallansch MA**. (2002) Molecular phylogeny and proposed classification of the simian picornaviruses. *Journal of Virology* 76(3):1244.
341. **Oberste MS, Maher K, Pallansch MA**. (2003) Genomic evidence that simian virus 2 and six other simian picornaviruses represent a new genus in Picornaviridae. *Virology* 314(1):283.
342. **Oberste MS, Maher K, Pallansch MA**. (2007) Complete genome sequences for nine simian enteroviruses. *Journal of General Virology* 88:3360.
343. **Olitsky PK**. (1945) Certain Properties of Theilers Virus, Especially in Relation to Its Use As Model for Poliomyelitis. *Proceedings of the Society for Experimental Biology and Medicine* 58(1):77.
344. **Ongus JR, Peters D, Bonmatin JM, Bengsch E, Vlak JM, van Oers MM**. (2004) Complete sequence of a picorna-like virus of the genus Iflavirus replicating in the mite Varroa destructor. *Journal of General Virology* 85:3747.
345. **Palmenberg AC**. (1989) Sequence alignments of picornaviral capsid proteins. Semler,B., Ehrenfeld,E. Molecular Aspects of Picornavirus Infection and Detection. 211. Washington, D. C., *American Society for Microbiology*.

346. **Palmenberg AC, Neubauer D, Skern T**. (2010) Genome Organization and Encoded Proteins. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. (1):3. Washington, *ASM Press*.
347. **Palmenberg AC, Pallansch MA, Rueckert RR**. (1979) Protease required for processing picornaviral coat protein resides in the viral replicase gene. *Journal of Virology* 32(3):770.
348. **Palmenberg AC, Parks GD, Hall DJ, Ingraham RH, Seng TW, Pallai PV**. (1992) Proteolytic processing of the cardiovirus P2 region: primary 2A/2B cleavage in clone-derived precursors. *Virology* 190(2):754.
349. **Palmenberg AC, Rathe JA, Liggett SB**. (2010) Analysis of the complete genome sequences of human rhinovirus. *Journal of Allergy and Clinical Immunology* 125(6):1190.
350. **Palmenberg AC, Sgro JY**. (2002) Alignments and comparative profiles of picornavirus genera. Semler,B., Wimmer,E. Molecular Biology of Picornaviruses. 149. Washington, D. C., *American Society for Microbiology*.
351. **Palmenberg AC, Spiro D, Kuzmickas R, Wang S, Djikeng A, Rathe JA, Fraser-Liggett CM, Liggett SB**. (2009) Sequencing and Analyses of All Known Human Rhinovirus Genomes Reveal Structure and Evolution. *Science* 324(5923):55.
352. **Pan JA, Peng XX, Gao YJ, Li ZL, Lu XL, Chen YZ, Ishaq M, Liu D, DeDiego ML, Enjuanes L et al.** (2008) Genome-Wide Analysis of Protein-Protein Interactions and Involvement of Viral Proteins in SARS-CoV Replication. *Plos One* 3(10):e3299.
353. **Pasternak AO, Spaan WJM, Snijder EJ**. (2006) Nidovirus transcription: how to make sense ... ? *Journal of General Virology* 87:1403.
354. **Paul AV, Molla A, Wimmer E**. (1994) Studies of a putative amphipathic helix in the N-terminus of poliovirus protein 2C. *Virology* 199(1):188.
355. **Paul AV, Schultz A, Pincus SE, Oroszlan S, Wimmer E**. (1987) Capsid Protein Vp4 of Poliovirus Is N-Myristoylated. *Proceedings of the National Academy of Sciences of the United States of America* 84(22):7827.
356. **Paul AV, van Boom JH, Filippov D, Wimmer E**. (1998) Protein-primed RNA synthesis by purified poliovirus RNA polymerase. *Nature* 393(6682):280.
357. **Pehrson JR, Fuji RN**. (1998) Evolutionary conservation of histone macroH2A subtypes and domains. *Nucl.Acids Res.* 26(12):2837.
358. **Pennisi E**. (2011) Going Viral: Exploring the Role of Viruses in our Bodies. *Science* 331:1513.
359. **Perl Foundation** (2011) The Perl programming language; http://www.perl.org
360. **Perlman S, Gallagher T, Snijder EJ**. (2008) Nidoviruses. Washington, DC, *ASM Press*.
361. **Pestova TV, Hellen CUT, Wimmer E**. (1994) A Conserved AUG Triplet in the 5' Nontranslated Region of Poliovirus Can Function As An Initiation Codon In-Vitro and In-Vivo. *Virology* 204(2):729.
362. **Petersen JFW, Cherney MM, Liebig HD, Skern T, Kuechler E, James MNG**. (1999) The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis. *EMBO J.* 18(20):5463.
363. **Pfister T, Jones KW, Wimmer E**. (2000) A cysteine-rich motif in poliovirus protein 2C(ATPase) is involved in RNA replication and binds zinc in vitro. *Journal of Virology* 74(1):334.
364. **Piccone ME, Chen HH, Roos RP, Grubman MJ**. (1996) Construction of a chimeric Theiler's murine encephalomyelitis virus containing the leader gene of foot-and-mouth disease virus. *Virology* 226(1):135.

365. **Piccone ME, Rieder E, Mason PW, Grubman MJ**. (1995) The Foot-And-Mouth-Disease Virus Leader Proteinase Gene Is Not Required for Viral Replication. *Journal of Virology* 69(9):5376.
366. **Plant EP, Perez-Alvarado GC, Jacobs JL, Mukhopadhyay B, Hennig M, Dinman JD**. (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *Plos Biology* 3(6):1012.
367. **Poch O, Sauvaget I, Delarue M, Tordo N**. (1989) Identification of 4 Conserved Motifs Among the RNA-Dependent Polymerase Encoding Elements. *Embo Journal* 8(12):3867.
368. **Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP**. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55(4):595.
369. **Porter FW, Palmenberg AC**. (2009) Leader-Induced Phosphorylation of Nucleoporins Correlates with Nuclear Trafficking Inhibition by Cardioviruses. *Journal of Virology* 83(4):1941.
370. **Posthuma CC, Nedialkova DD, Zevenhoven-Dobbe JC, Blokhuis JH, Gorbalenya AE, Snijder EJ**. (2006) Site-directed mutagenesis of the nidovirus replicative endoribonuclease NendoU exerts pleiotropic effects on the arterivirus life cycle. *Journal of Virology* 80(4):1653.
371. **Prentice E, McAuliffe J, Lu XT, Subbarao K, Denison MR**. (2004) Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. *Journal of Virology* 78(18):9977.
372. **Price AL, Jones NC, Pevzner PA**. (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21:I351-I358.
373. **Pringle CR**. (1991) The 20Th Meeting of the Executive-Committee of the International-Committee-On-Virus-Taxonomy - Virus Species, Higher Taxa, A Universal Virus Database, and Other Matters. *Archives of Virology* 119(3-4):303.
374. **Pruitt KD, Tatusova T, Klimke W, Maglott DR**. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research* 37:D32-D36.
375. **Putics A, Filipowicz W, Hall J, Gorbalenya AE, Ziebuhr J**. (2005) ADP-ribose-1 "-monophosphatase: a conserved coronavirus enzyme that is dispensable for viral replication in tissue culture. *Journal of Virology* 79(20):12721.
376. **Qi XX, Lan SY, Wang WJ, Schelde LM, Dong HH, Wallat GD, Ly H, Liang YY, Dong CJ**. (2010) Cap binding and immune evasion revealed by Lassa nucleoprotein structure. *Nature* 468(7325):779-U65.
377. **R Development Core Team** (2011) R: A Language and Environment for Statistical Computing; http://www.R-project.org
378. **Racaniello V**. (2007) Picornaviridae: the viruses and their replication. Knipe,D. et al. Fields Virology. 5th(24):795. Philadlephia, *Wolters Kluwer*.
379. **Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H**. (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B-Biological Sciences* 275(1632):237.
380. **Ramsay JO**. (1988) Monotone Regression Splines in Action. *Statistical Science* 3(4):425. *Institute of Mathematical Statistics*.
381. **Rannala B, Yang ZH**. (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* 43(3):304.
382. **Raup DM**. (1994) The Role of Extinction in Evolution. *Proceedings of the National Academy of Sciences of the United States of America* 91(15):6758.

383. **Reavy B, Mayo MA, Turnbullross AD, Murant AF**. (1993) Parsnip Yellow Fleck and Rice Tungro Spherical Viruses Resemble Picornaviruses and Represent 2 Genera in A Proposed New Plant Picornavirus Family (Sequiviridae). *Archives of Virology* 131(3-4):441.
384. **Reddick BB, Habera LF, Law MD**. (1997) Nucleotide sequence and taxonomy of maize chlorotic dwarf virus within the family Sequiviridae. *Journal of General Virology* 78:1165.
385. **Reeder J, Steffen P, Giegerich R**. (2007) pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research* 35:W320-W324.
386. **Reuter G, Boldizsar A, Pankovics P**. (2009) Complete nucleotide and amino acid sequences and genetic organization of porcine kobuvirus, a member of a new species in the genus Kobuvirus, family Picornaviridae. *Archives of Virology* 154(1):101.
387. **Ricagno S, Egloff MP, Ulfertst R, Coutard B, Nurizzo D, Campanacci V, Cambillau C, Ziebuhrt J, Canard B**. (2006) Crystal structure and mechanistic determinants of SARS coronavirus nonstructural protein 15 define an endoribonuclease family. *Proceedings of the National Academy of Sciences of the United States of America* 103(32):11892.
388. **Ricour C, Borghese F, Sorgeloos F, Hato SV, van Kuppeveld FJM, Michiels T**. (2009) Random Mutagenesis Defines a Domain of Theiler's Virus Leader Protein That Is Essential for Antagonism of Nucleocytoplasmic Trafficking and Cytokine Gene Expression. *Journal of Virology* 83(21):11223.
389. **Rietveld K, Vanpoelgeest R, Pleij CWA, Vanboom JH, Bosch L**. (1982) The Transfer Rna-Like Structure at the 3' Terminus of Turnip Yellow Mosaic-Virus Rna - Differences and Similarities with Canonical Transfer-Rna. *Nucleic Acids Research* 10(6):1929.
390. **Riquet FB, Blanchard C, Jegouic S, Balanant J, Guillot S, Vibet MA, Rakoto-Andrianarivelo M, Delpeyroux F**. (2008) Impact of exogenous sequences on the characteristics of an epidemic type 2 recombinant vaccine-derived poliovirus. *Journal of Virology* 82(17):8927.
391. **Rodriguez PL, Carrasco L**. (1993) Poliovirus protein 2C has ATPase and GTPase activities. *Journal of Biological Chemistry* **268**:8105.
392. **Romanova LI, Lidsky PV, Kolesnikova MS, Fominykh KV, Gmyl AP, Sheval EV, Hato SV, van Kuppeveld FJM, Agol VI**. (2009) Antiapoptotic Activity of the Cardiovirus Leader Protein, a Viral "Security" Protein. *Journal of Virology* 83(14):7273.
393. **Rossmann MG, Arnold E, Erickson JW, Frankenberger EA, Griffith JP, Hecht H-J, Johnson JE, Kamer G, Luo M, Mosser AG et al.** (1985) Structure of human cold virus and functional relationship to other picornaviruses. *Nature (London)* 317:145.
394. **Rossmann MG, Johnson JE**. (1989) Icosahedral RNA Virus Structure. *Ann.Rev.Biochem.* 58:533.
395. **Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al.** (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348.
396. **Rozanov MN, Koonin EV, Gorbalenya AE**. (1992) Conservation of the Putative Methyltransferase Domain - A Hallmark of the Sindbis-Like Supergroup of Positive-Strand Rna Viruses. *Journal of General Virology* 73:2129.

397. **Rozovics JM, Semler B**. (2010) Genome Replication I: the Players. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 107. Washington, DC, *ASM Press*.
398. **Rueckert RR, Wimmer E**. (1984) Systematic nomenclature of picornavirus proteins. *Journal of Virology* 50(3):957.
399. **Rux JJ, Burnett RM**. (1998) Spherical viruses. *Curr.Opin.Struct.Biol.* 8(2):142.
400. **Ryabov EV**. (2007) A novel virus isolated from the aphid Brevicoryne brassicae with similarity to Hymenoptera picorna-like viruses. *Journal of General Virology* 88:2590.
401. **Ryan MD, Flint M**. (1997) Virus-encoded proteinases of the picornavirus super-group. *Journal of General Virology* 78(Pt 4):699.
402. **Saikatendu KS, Joseph JS, Subramanian V, Clayton T, Griffith M, Moy K, Velasquez J, Neuman BW, Buchmeier MJ, Stevens RC et al.** (2005) Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1 "-phosphate dephosphorylation by a conserved domain of nsP3. *Structure* 13(11):1665.
403. **Sanchez G, Bosch A, Gomez-Mariano G, Domingo E, Pinto RM**. (2003) Evidence for quasispecies distributions in the human hepatitis A virus genome. *Virology* 315(1):34.
404. **Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R**. (2010) Viral Mutation Rates. *Journal of Virology* 84(19):9733.
405. **Santti J, Hyypia T, Kinnunen L, Salminen M**. (1999) Evidence of recombination among enteroviruses. *Journal of Virology* 73(10):8741.
406. **Sarkar IN, Thornton JW, Planet PJ, Figurski DH, Schierwater B, DeSalle R**. (2002) An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution* 24(3):388.
407. **Savolainen C, Blomqvist S, Mulders MN, Hovi T**. (2002) Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. *Journal of General Virology* 83:333.
408. **Sawicki SG, Sawicki DL, Siddell SG**. (2007) A contemporary view of coronavirus transcription. *Journal of Virology* 81(1):20.
409. **Sawicki SG, Sawicki DL, Younker D, Meyer Y, Thiel V, Stokes H, Siddell SG**. (2005) Functional and genetic analysis of coronavirus replicase-transcriptase proteins. *Plos Pathogens* 1(4):310.
410. **Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S et al.** (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 38:D5-D16.
411. **Schein CH, Oezguen N, Volk DE, Garimella R, Paul A, Braun W**. (2006) NMR structure of the viral peptide linked to the genome (VPg) of poliovirus. *Peptides* 27(7):1676.
412. **Schmidt HA, Strimmer K, Vingron M, von Haeseler A**. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3):502.
413. **Schuffenecker I, Ando T, Thouvenot D, Lina B, Aymard M**. (2001) Genetic classification of "Sapporo-like viruses". *Archives of Virology* 146(11):2115.
414. **Schutze H, Ulferts R, Schelle B, Bayer S, Granzow H, Hoffmann B, Mettenleiter TC, Ziebuhr J**. (2006) Characterization of White bream virus reveals a novel genetic cluster of nidoviruses. *Journal of Virology* 80(23):11598.
415. **Semler BL, Wimmer E**. (2002) Molecular biology of picornaviruses. Semler,B.L., Wimmer,E. Washington, DC, U.S.A., *ASM Press*.

416. **Sepkoski JJ**. (1998) Rates of speciation in the fossil record. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 353(1366):315.
417. **Seybert A, Hegyi A, Siddell SG, Ziebuhr J**. (2000) The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5'-to-3' polarity. *RNA* 6(7):1056.
418. **Seybert A, Posthuma CC, van Dinten LC, Snijder EJ, Gorbalenya AE, Ziebuhr J**. (2005) A complex zinc finger controls the enzymatic activities of nidovirus helicases. *Journal of Virology* 79(2):696.
419. **Seybert A, van Dinten LC, Snijder EJ, Ziebuhr J**. (2000) Biochemical characterization of the equine arteritis virus helicase suggests a close functional relationship between arterivirus and coronavirus helicases. *Journal of Virology* 74(20):9586.
420. **Shapiro B, Rambaut A, Pybus OG, Holmes EC**. (2006) A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Molecular Biology and Evolution* 23(9):1724.
421. **Shi ST, Schiller JJ, Kanjanahaluethai A, Baker SC, Oh JW, Lai MM**. (1999) Colocalization and membrane association of murine hepatitis virus gene 1 products and De novo-synthesized viral RNA in infected cells. *Journal of Virology* 73(7):5957.
422. **Shukla DD, Ward CW**. (1988) Amino-Acid Sequence Homology of Coat Proteins As A Basis for Identification and Classification of the Potyvirus Group. *Journal of General Virology* 69:2703.
423. **Sidorov IA, Samborskiy DV, Leontovich AM, Gorbalenya AE** (2012) HAYGENS, Homology-Annotation hYbrid retrieval of GENetic Sequences; http://veb.lumc.nl/HAYGENS/index.cgi
424. **Siew N, Azaria Y, Fischer D**. (2004) The ORFanage: an ORFan database. *Nucl.Acids Res.* 32:D281-D283.
425. **Simmonds P**. (2006) Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses. *Journal of Virology* 80(22):11124.
426. **Simmonds P**. (2010) Recombination in the Evolution of Picornaviruses. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 229. Washington, DC, *ASM Press*.
427. **Simmonds P, McIntyre C, Savolainen-Kopra C, Tapparel C, Mackay IM, Hovi T**. (2010) Proposals for the classification of human rhinovirus species C into genotypically assigned types. *Journal of General Virology* 91:2409.
428. **Simmonds P, Welch J**. (2006) Frequency and dynamics of recombination within different species of human enteroviruses. *Journal of Virology* 80(1):483.
429. **Sittidjlokratna N, Dangtip S, Cowley JA, Walker PJ**. (2008) RNA transcription analysis and completion of the genome sequence of yellow head nidovirus. *Virus Research* 136(1-2):157.
430. **Skern T, Hampoelz B, Guarne A, Fita I, Bergmann E, Petersen J, James MNG**. (2002) Structure and Function of Picornavirus Proteinases. Semler,B.L., Wimmer,E. Molecular Biology of Picornaviruses. (17):199. Washington, D. C., *American Society for Microbiology*.
431. **Sniegowski PD, Gerrish PJ, Johnson T, Shaver A**. (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* 22(12):1057.
432. **Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE**. (2003) Unique and conserved features

of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *Journal of Molecular Biology* 331(5):991.

433. **Snijder EJ, Wassenaar ALM, vanDinten LC, Spaan WJM, Gorbalenya AE**. (1996) The arterivirus Nsp4 protease is the prototype of a novel group of chymotrypsin-like enzymes, the 3C-like serine proteases. *Journal of Biological Chemistry* 271(9):4864.
434. **Soding J**. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951.
435. **Stanway G, Brown F, Christian P, Hovi T, Hyypiae T, King AMQ, Knowles NJ, Lemon SM, Minor PD, Pallansch MA et al.** (2005) Family *Picornaviridae*. Fauquet,C.M. et al. Virus Taxonomy, Eighth report of the International Committee on Taxonomy of Viruses. 757. *Elsevier Academic Press*.
436. **Stanway G, Hovi T, Knowles NJ, Hyypia T**. (2002) Molecular and Biological Basis of Picornavirus Taxonomy. Semler,B.L., Wimmer,E. Molecular Biology of Picornaviruses. 17. Washington, DC, *Am Soc Microbiol Press*.
437. **Stanway G, Hyypia T**. (1999) Parechoviruses. *Journal of Virology* 73(7):5249.
438. **Steil BP, Barton DJ**. (2009) Cis-active RNA elements (CREs) and picornavirus RNA replication. *Virus Research* 139(2):240.
439. **Steinhauer DA, Domingo E, Holland JJ**. (1992) Lack of Evidence for Proofreading Mechanisms Associated with An Rna Virus Polymerase. *Gene* 122(2):281.
440. **Summers DF, Maizel JV, Darnell JE**. (1965) Evidence for Virus-Specific Noncapsid Proteins in Poliovirus-Infected Hela Cells. *Proceedings of the National Academy of Sciences of the United States of America* 54(2):505.
441. **Sweeney TR, Dhote V, Yu YP, Hellen CUT**. (2012) A Distinct Class of Internal Ribosomal Entry Site in Members of the Kobuvirus and Proposed Salivirus and Paraturdivirus Genera of the Picornaviridae. *Journal of Virology* 86(3):1468.
442. **Szathmary E, Smith JM**. (1995) The Major Evolutionary Transitions. *Nature* 374(6519):227.
443. **Talavera G, Castresana J**. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56(4):564.
444. **Tate J, Liljas L, Scotti P, Christian P, Lin TW, Johnson JE**. (1999) The crystal structure of cricket paralysis virus: the first view of a new virus family. *Nature Structural Biology* 6(8):765.
445. **te Velthuis AJW, Arnold JJ, Cameron CE, van den Worm SHE, Snijder EJ**. (2010) The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. *Nucleic Acids Research* 38(1):203.
446. **Teterina NL, Gorbalenya AE, Egger D, Bienz K, Ehrenfeld E**. (1997) Poliovirus 2C protein determinants of membrane binding and rearrangements in mammalian cells. *Journal of Virology* 71(12):8962.
447. **Teterina NL, Gorbalenya AE, Egger D, Bienz K, Rinaudo MS, Ehrenfeld E**. (2006) Testing the modularity of the N-terminal amphipathic helix conserved in picornavirus 2C proteins and hepatitis C NS5A protein. *Virology* 344(2):453.
448. **Teterina NL, Lauber C, Jensen KS, Levenson EA, Gorbalenya AE, Ehrenfeld E**. (2011) Identification of tolerated insertion sites in poliovirus non-structural proteins. *Virology* 409(1):1.
449. **Theiler M**. (1934) Spontaneous encephalomyelitis of mice - a new virus disease. *Science* 80(2066):122.

450. **Thiel V, Herold J, Schelle B, Siddell SG**. (2001) Viral replicase gene products suffice for coronavirus discontinuous transcription. *Journal of Virology* 75(14):6676.
451. **Thompson AA, Peersen OB**. (2004) Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *EMBO J.* 23(17):3462.
452. **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 Research* 22(22):4673.
453. **Tibayrenc M**. (2006) The species concept in parasites and other pathogens: a pragmatic approach? *Trends in Parasitology* 22(2):66.
454. **Tijms MA, Nedialkova DD, Zevenhoven-Dobbe JC, Gorbalenya AE, Snijder EJ**. (2007) Arterivirus subgenomic rnRNA synthesis and virion biogenesis depend on the multifunctional nsp1 autoprotease. *Journal of Virology* 81(19):10496.
455. **Tijms MA, van Dinten LC, Gorbalenya AE, Snijder EJ**. (2001) A zinc finger-containing papain-like protease couples subgenomic mRNA synthesis to genome translation in a positive-stranded RNA virus. *Proceedings of the National Academy of Sciences of the United States of America* 98(4):1889.
456. **Tolskaya EA, Romanova LA, Kolesnikova MS, Agol VI**. (1983) Intertypic Recombination in Poliovirus - Genetic and Biochemical-Studies. *Virology* 124(1):121.
457. **Tolskaya EA, Romanova LI, Kolesnikova MS, Gmyl AP, Gorbalenya AE, Agol VI**. (1994) Genetic studies on the poliovirus 2C protein, an NTPase. A plausible mechanism of guanidine effect on the 2C function and evidence for the importance of 2C oligomerization. *J.Mol.Biol.* 236(5):1310.
458. **Toyoda H, Nicklin MJH, Murray MG, Anderson CW, Dunn JJ, Studier FW, Wimmer E**. (1986) A 2Nd Virus-Encoded Proteinase Involved in Proteolytic Processing of Poliovirus Polyprotein. *Cell* 45(5):761.
459. **Tseng CH, Knowles NJ, Tsai HJ**. (2007) Molecular analysis of duck hepatitis virus type 1 indicates that it should be assigned to a new genus. *Virus Research* 123(2):190.
460. **Tseng CH, Tsai HJ**. (2007) Sequence analysis of a duck picornavirus isolate indicates that it together with porcine enterovirus type 8 and simian picornavirus type 2 should be assigned to a new picornavirus genus. *Virus Research* 129(2):104.
461. **van Dinten LC, van Tol H, Gorbalenya AE, Snijder EJ**. (2000) The predicted metal-binding region of the arterivirus helicase protein is involved in subgenomic mRNA synthesis, genome replication, and virion biogenesis. *Journal of Virology* 74(11):5213.
462. **van Hemert MJ, van den Worm SHE, Knoops K, Mommaas AM, Gorbalenya AE, Snijder EJ**. (2008) SARS-coronavirus replication/transcription complexes are membrane-protected and need a host factor for activity in vitro. *Plos Pathogens* 4(5):e1000054.
463. **van Kuppeveld FJM, Belov G, Ehrenfeld E**. (2010) Remodeling Cellular Membranes. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 181. Washington, DC, *ASM Press*.
464. **van Ooij MJM, Vogt DA, Paul A, Castro C, Kuijpers J, van Kuppeveld FJM, Cameron CE, Wimmer E, Andino R, Melchers WJG**. (2006) Structural and functional characterization of the coxsackievirus B3 CRE(2C): role of CRE(2C) in negative- and positive-strand RNA synthesis. *Journal of General Virology* 87:103.

465. **Van Regenmortel MHV**. (1989) Applying the Species Concept to Plant-Viruses. *Archives of Virology* 104(1-2):1.
466. **Van Regenmortel MHV**. (2003) Viruses are real, virus species are man-made, taxonomic constructions. *Archives of Virology* 148(12):2481.
467. **Van Regenmortel MHV**. (2007) Virus species and virus identification: Past and current controversies. *Infection Genetics and Evolution* 7(1):133.
468. **Venkataraman S, Reddy SP, Loo J, Idamakanti N, Hallenbeck PL, Reddy VS**. (2008) Structure of Seneca Valley Virus-001: An Oncolytic Picornavirus Representing a New Genus. *Structure* 16(10):1555.
469. **Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL**. (2008) Rapid identification of known and new RNA viruses from animal tissues. *Plos Pathogens* 4(9).
470. **Vignuzzi M, Andino R**. (2010) Biological Implications of Picornavirus Fidelity Mutants. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. 213. Washington, DC, *ASM Press*.
471. **Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R**. (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439(7074):344.
472. **Vingron M, Argos P**. (1989) A Fast and Sensitive Multiple Sequence Alignment Algorithm. *Computer Applications in the Biosciences* 5(2):115.
473. **Ward CD, Flanegan JB**. (1992) Determination of the Poliovirus Rna-Polymerase Error Frequency at 8 Sites in the Viral Genome. *Journal of Virology* 66(6):3784.
474. **Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ**. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189.
475. **Welsh J, Mcclelland M**. (1990) Fingerprinting Genomes Using PCR with Arbitrary Primers. *Nucleic Acids Research* 18(24):7213.
476. **Wesche PL, Gaffney DJ, Keightley PD**. (2004) DNA sequence error rates in Genbank records estimated using the mouse genome as a reference. *Dna Sequence* 15(5-6):362.
477. **Wessels E, Notebaart RA, Duijsings D, Lanke K, Vergeer B, Melchers WJG, van Kuppeveld FJM**. (2006) Structure-function analysis of the coxsackievirus protein 3A - Identification of residues important for dimerization, viral RNA replication, and transport inhibition. *Journal of Biological Chemistry* 281(38):28232.
478. **Whelan S, Goldman N**. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* 18(5):691.
479. **Williams CH, Kajander T, Hyypia T, Jackson T, Sheppard D, Stanway G**. (2004) Integrin alpha(v)beta(6) is an RGD-dependent receptor for coxsackievirus A9. *Journal of Virology* 78(13):6967.
480. **Wimmer E, Hellen CUT, Cao XM**. (1993) Genetics of Poliovirus. *Annual Review of Genetics* 27:353.
481. **Wimmer E, Paul A**. (2010) Making of a picornavirus genome. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. (3):33. Washington, *ASM Press*.
482. **Wimmer E, Paul A**. (2010) The Making of a picornavirus genome. Ehrenfeld,E., Domingo,E., Roos,R.P. The Picornaviruses. (3):33. Washington, *ASM Press*.
483. **Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Bocker S, Stoye J, Baumbach J**. (2010) Partitioning biological data with transitivity clustering. *Nature Methods* 7(6):419.

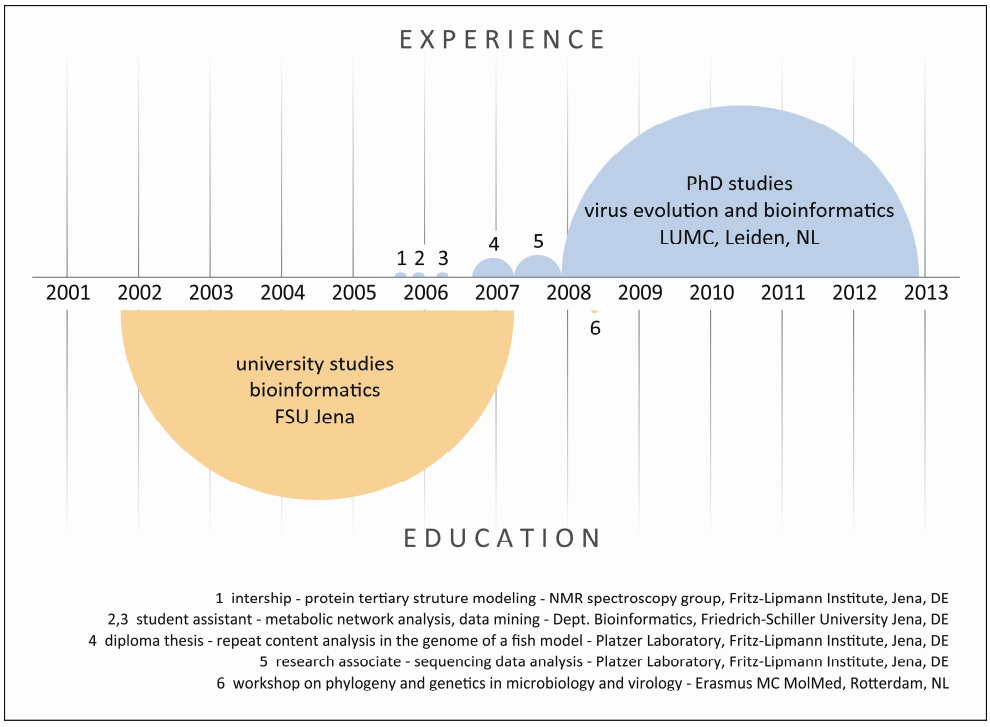
484. **Wutz G, Auer H, Nowotny N, Grosse B, Skern T, Kuechler E**. (1996) Equine rhinovirus serotypes 1 and 2: Relationship to each other and to aphthoviruses and cardioviruses. *Journal of General Virology* 77:1719.
485. **Yang Y, Rijnbrand R, McKnight KL, Wimmer E, Paul A, Martin A, Lemon SM**. (2002) Sequence requirements for viral RNA replication and VPg uridylylation directed by the internal cis-acting replication element (cre) of human rhinovirus type 14. *Journal of Virology* 76(15):7485.
486. **Yang Y, Yi MK, Evans DJ, Simmonds P, Lemon SM**. (2008) Identification of a Conserved RNA Replication Element (cre) within the 3D(pol)-Coding Sequence of Hepatoviruses. *Journal of Virology* 82(20):10118.
487. **Yang ZH**. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24(8):1586.
488. **Yu SF, Lloyd RE**. (1992) Characterization of the roles of conserved cysteine and histidine residues in poliovirus 2A protease. *Virology* **186**:725.
489. **Zanotto PMD, Gibbs MJ, Gould EA, Holmes EC**. (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *Journal of Virology* 70(9):6083.
490. **Zeddam JL, Gordon KHJ, Lauber C, Alves CAF, Luke BT, Hanzlik TN, Ward VK, Gorbalenya AE**. (2010) Euprosterna elaeasa virus genome sequence and evolution of the Tetraviridae family: Emergence of bipartite genomes and conservation of the VPg signal with the dsRNA Birnaviridae family. *Virology* 397(1):145.
491. **Zhai YJ, Sun F, Li XM, Pang H, Xu XL, Bartlam M, Rao ZH**. (2005) Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer. *Nat.Struct.Mol.Biol.* 12(11):980.
492. **Zhang JY, Temin HM**. (1994) Retrovirus Recombination Depends on the Length of Sequence Identity and Is Not Error-Prone. *Journal of Virology* 68(4):2409.
493. **Zhang L, Sato S, Kim JI, Roos RP**. (1995) Theilers Virus As A Vector for Foreign Gene Delivery. *Journal of Virology* 69(5):3171.
494. **Zhao L, Jha BK, Wu A, Elliott R, Ziebuhr J, Gorbalenya AE, Silverman RH, Weiss SR**. (2012) Antagonism of the Interferon-Induced OAS-RNase L Pathway by Murine Coronavirus ns2 Protein Is Required for Virus Replication and Liver Pathology. *Cell Host & Microbe* 11(6):607.
495. **Zhao WD, Wimmer E, Lahser FC**. (1999) Poliovirus/hepatitis C virus (internal ribosomal entry site-core) chimeric viruses: Improved growth properties through modification of a proteolytic cleavage site and requirement for core RNA sequences but not for core-related polypeptides. *Journal of Virology* 73(2):1546.
496. **Zheng DP, Ando T, Fankhauser RL, Beard RS, Glass RI, Monroe SS**. (2006) Norovirus classification and proposed strain nomenclature. *Virology* 346(2):312.
497. **Zheng T**. (2007) Characterisation of two enteroviruses isolated from Australian brushtail possums (Trichosurus vulpecula) in New Zealand. *Archives of Virology* 152(1):191.
498. **Ziebuhr J, Snijder EJ, Gorbalenya AE**. (2000) Virus-encoded proteinases and proteolytic processing in the Nidovirales. *Journal of General Virology* 81:853.
499. **Zimmern D**. (1988) Evolution of RNA Viruses. Domingo,E., Holland,J.J., Ahlquist,P. RNA Genetics. 211. Boca Raton, FL, *CRC Press*.
500. **Zirkel F, Kurth A, Quan PL, Briese T, Ellerbrok H, Pauli G, Leendertz FH, Lipkin WI, Ziebuhr J, Drosten C et al.** (2011) An Insect Nidovirus Emerging from a Primary Tropical Rainforest. *mBio* 2(3):e00077.

501. **Zoll J, Galama JMD, van Kuppeveld FJM**. (2009) Identification of Potential Recombination Breakpoints in Human Parechoviruses. *Journal of Virology* 83(7):3379.
502. **Zuker M**. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* 31(13):3406.
503. **Zust R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, Ziebuhr J, Szretter KJ, Baker SC, Barchet W, Diamond MS et al.** (2011) Ribose 2 '-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nature Immunology* 12(2):137-U46.

Acknowledgments

This thesis would not have been possible without the help of many people. I would like to thank my promotor Sasha for giving me the opportunity to do my PhD studies in his group, for his continuous support, and for all the advice he has given me. I have learned so many things from you! I am grateful to my colleagues at the Department of Medical Microbiology for a nice working atmosphere and helpful discussions and suggestions on my rather exotic bioinformatics projects. Special thanks go to the former head of department Willy and the current head of department Louis for supporting bioinformatics in this rather medically oriented department. I also would like to thank Corrie, Ria, Manon, and Dick for their help with administrative issues. I am thankful to Igor, Kalina, and Misha for being great office mates, for numerous helpful suggestions, and for nice conversations. I also would like to thank our system administrator Hans for technical support as well as Andrey, Alex, and Dima in Moscow for maintaining and advancing the Viralis software that was essential to performing my studies. I am grateful to all my co-authors, especially to Christian, Jelle, John, Eric, Maria, Kouichi, and Sandra for excellent comments on manuscripts and the great experience to work with you. On a more personal side, I would like to thank my family and friends for their unconditional support and the countless nice moments that make life so enjoyable. And I thank my girlfriend Katrin for always standing by me, for accepting my weaknesses, and for the love she has given me. What would I do without you!?

Curriculum vitae



Chris Lauber was born on September 15, 1981 in Gera, Germany. He obtained his *university-entrance diploma* (Abitur) in 2000 from the Georg-Christoph Lichtenberg Gymnasium, Gera, Germany. His studies in bioinformatics at the Friedrich-Schiller-Universität Jena, Germany from 2001 to 2007 resulted in the *master thesis* (Diplom) “Initial characterization of repetitive elements in the genome of *Nothobranchius furzeri*” supervised by Dr. Kathrin Reichwald and Dr. Matthias Platzer at the Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany. He performed his PhD studies in virus evolution and bioinformatics under the supervision of Prof. Dr. Alexander E. Gorbalenya at the Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands between 2007 and 2012, which converged to this thesis.

Other publications

Lauber C, Gorbalenya AE (2012) Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses* 4:1425.

Knetsch CW, Terveer EM, **Lauber C**, Gorbalenya AE, Harmanus C, Kuijper EJ, Corver J, van Leeuwen HC (2012) Comparative analysis of an expanded *Clostridium difficile* reference strain collection reveals genetic diversity and evolution through six lineages. *Infection, Genetics and Evolution* 12:1577.

Dorrington RA, Gorbalenya AE, Gordon KHJ, **Lauber C**, Ward VK (2011) Family Tetraviridae. In: King A, Adams M, Carstens E, Lefkowitz EJ. Virus Taxonomy - Ninth Report of the International Committee on Taxonomy of Viruses. *Academic Press*, pp. 1091-1102.

Zlateva KT, Crusio KM, Leontovich AM, **Lauber C**, Claas E, Kravchenko A, Spaan WJM, Gorbalenya AE (2011) Design and validation of consensus-degenerate hybrid oligonucleotide primers for broad and sensitive detection of corona- and toroviruses. *Journal of Virological Methods* 177:174.

Teterina NL, **Lauber C**, Jensen KS, Levenson EA, Gorbalenya AE, Ehrenfeld E (2011) Identification of tolerated insertion sites in poliovirus non-structural proteins. *Virology* 409:1.

Van der Meijden E, Janssens RWA, **Lauber C**, Bouwes Bavinck JN, Gorbalenya AE, Feltkamp MCW (2010) Discovery of a new Human Polyomavirus associated with Trichodysplasia Spinulosa in an immunocompromized patient. *PLoS Pathogens* 6:e1001024.

Assenberg R, Delmas O, Morin B, Graham SC, De Lamballerie X, **Lauber C**, Coutard B, Grimes JM, Neyts J, Owens RJ, Brandt BW, Gorbalenya AE, Tucker P, Stuart DI, Canard B, Bourhy H (2010) Genomics and structure/function studies of *Rhabdoviridae* proteins involved in replication and transcription. *Antiviral Research* 87:149.

Zeddam JL, Gordon KHJ, **Lauber C**, Felipe Alves CA, Luke BT, Hanzlik TN, Ward VK, Gorbalenya AE (2010) Euprosterna elaeasa virus genome sequence and evolution of the *Tetraviridae* family: Emergence of bipartite genome and convervation of the VPg signal with the dsRNA *Birnaviridae* family. *Virology* 397:145.

Nedialkova DD, Ulferts R, van den Born E, **Lauber C**, Gorbalenya AE, Ziebuhr J, Snijder EJ (2009) Biochemical characterization of Arterivirus nonstructural protein 11 reveals the Nidovirus-wide conservation of a replicative endoribonuclease. *Journal of Virology* 83:5671.

Reichwald K, **Lauber C**, Nanda I, Kirschner J, Hartmann N, Schories S, Gausmann U, Taudien S, Schilhabel MB, Szafranski K, Glöckner G, Schmidt M, Cellerino A, Schartl M, Englert C, Platzer M (2009) High tandem repeat content in the genome of the short-lived anuual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biology* 10: R16.