

**Part 1:** **TITLE, AUTHORS, APPROVALS, etc**

|  |  |  |
| --- | --- | --- |
| **Code assigned:** | **2021.022B** |  |
| **Short title:** Create one new order (*Crassvirales*) including four new families, ten new subfamilies, 42 new genera and 73 new species (*Caudoviricetes*) | | |
|  | | |

**Author(s) and email address(es)**

|  |  |
| --- | --- |
| Shkoporov AN, Stockdale SR, Adriaenssens EM, Yutin N, Koonin EV, Dutilh BE, Krupovic M, Edwards RA, Tolstoy I, Hill C | [andrey.shkoporov@ucc.ie](mailto:andrey.shkoporov@ucc.ie)  [stephen.stockdale@ucc.ie](mailto:stephen.stockdale@ucc.ie)  evelien.adriaenssens@quadram.ac.uk  [yutin@ncbi.nlm.nih.gov](mailto:yutin@ncbi.nlm.nih.gov)  [koonin@ncbi.nlm.nih.gov](mailto:koonin@ncbi.nlm.nih.gov)  [bedutilh@gmail.com](mailto:bedutilh@gmail.com)  [mart.krupovic@pasteur.fr](mailto:mart.krupovic@pasteur.fr)  robert.edwards@flinders.edu.au  [tolstoy@ncbi.nlm.nih.gov](mailto:tolstoy@ncbi.nlm.nih.gov)  [c.hill@ucc.ie](mailto:c.hill@ucc.ie) |

**Author(s) institutional address(es) (optional)**

|  |
| --- |
| University College Cork, Ireland [ANS, SRS, CH]  NCBI, NLM, NIH, USA [YN, EVK, IT]  Quadram Institute, UK [EMA]  Utrecht University, The Netherlands [BED]  Flinders University, Australia [RAE] |

**Corresponding author**

|  |
| --- |
| Andrey N Shkoporov, Chair of ICTV crAss-like phages Study Group |

**List the ICTV Study Group(s) that have seen this proposal**

|  |  |
| --- | --- |
| crAss-like phages Study Group, Bacterial Viruses Subcommittee |  |

**ICTV study group comments and response of proposer**

|  |
| --- |
| NA |

**Authority to use the name of a living person**

|  |  |
| --- | --- |
| **Is any taxon name used here derived from that of a living person (Y/N)** | N |

|  |  |  |
| --- | --- | --- |
| **Taxon name** | **Person from whom the name is derived** | **Permission attached (Y/N)** |
|  |  |  |
|  |  |  |
|  |  |  |

**Submission dates**

|  |  |
| --- | --- |
| Date first submitted to SC Chair | 27/7/2020 |
| Date of this revision (if different to above) | 13/5/2021 |

**ICTV-EC comments and response of the proposer**

|  |
| --- |
| This is a resubmission of proposal 2020.039B. The proposal was submitted to the EC by the Bacterial and Archaeal Viruses Subcommittee Chair and discussed at EC52. It was given provisional approval based on the evidence presented, to be approved by email vote. The minor revisions requested by the EC were that all exemplar isolate genomes were to be publicly available from GenBank as complete and annotated sequence records.  Not all accession numbers were assigned in time for the email vote, as a result the proposal was deferred to EC53.  In this revised proposal, new exemplar genomes were chosen and all now have accession numbers associated with them. As a result, two of the previously proposed families do not have genomes deposited in the appropriate databases, and we have suspended their creation for the time being. |

**Part 3:** **TAXONOMIC PROPOSAL**

**Name of accompanying Excel module**

|  |
| --- |
| 2021.022B.R.Crassvirales |

**Abstract**

|  |
| --- |
| CrAss-like phages are a diverse group of mostly uncultured bacterial viruses that are highly abundant in the mammalian gut and other habitats. First identified in metagenomic sequences from human faeces in 2014, crAss-like phages were predicted to infect members of the phylum Bacteroidetes. Later work resulted in the isolation of the first cultured representatives, the confirmation of a *Podoviridae*-like morphology and a proposal to classify uncultured crAss-like phages under a novel taxonomic group. Based on substantial genomic differences between this group and other members of the order *Caudovirales*, we propose the creation of a new order *Crassvirales*, comprising at the moment four new families, ten new subfamilies, 42 new genera and a total of 73 new species. Phylogenetic trees of conserved virus genes, supported by the levels of gene sharing, were used to delineate novel families, subfamilies and genera, and levels of average nucleotide identity were used for species demarcation. |

**Text of proposal**

|  |  |
| --- | --- |
| |  | | --- | | **Family demarcation criteria:** Phylogenies of highly conserved proteins (large terminase subunit, major capsid protein, DNA primase) were employed as the principal family demarcation criterion. Each of the 6 proposed families in the new order (*Intestiviridae*, *Crevaviridae*, *Suoliviridae*, *Steigviridae*) is a strongly supported clade in these trees. The tree topology is supported by analysis of shared orthologous genes (as determined using OrthoMCL), with the families sharing at least 17% of the genes.  **Subfamily demarcation criteria:** Subfamilies were inferred from protein phylogenetic trees as the major second order branches; the boundaries between them in terms of percentage of shared genes were less apparent than the inter-family boundaries. A minimum of 27-79% of genes are shared between member species depending on a particular subfamily.  **Genus demarcation criteria:** The genera within *Crassvirales* were defined based on the topology of the protein phylogenetic trees which was supported by sharing at least 80% of orthologous genes.  **Species demarcation criteria:** Based on the recommended MIUViG standard [1], we chose 95% DNA sequence identity over 85% of complete genome, or circular metagenome-assembled genome (cMAG) length as the criterion for demarcation of species. Each of the proposed species differs from the others by more than 5% at the DNA level as confirmed by BLASTn search implemented in PyANI software. | |

**Supporting evidence**

**Origins of the names of the taxa:**

The proposed new order *Crassvirales* is named after the prototypical uncultured virus, crAssphage, which is highly abundant in the human gut, but was overlooked until its discovery in 2014 by Dutilh and colleagues [2]. The name of this uncultured virus is based on the software tool crAss for cross-assembly analysis [3] that was used for the discovery of the genome in multiple faecal metagenomic sequencing datasets [4].

Because crAss-like phages (phages similar in their genomic architecture and sharing the same ancestor with crAssphage, based on conserved protein phylogenies) are highly abundant in mammalian gut, particularly, in humans [5-8], the names of the four proposed families derive from the word ‘intestines’ in four languages: *Intestiviridae* (Latin), *Crevaviridae* (*creva*, Bosnian), *Suoliviridae* (*suolet*, Finnish), *Steigviridae* (*stéig*, Irish).

The ten rough synonyms of the word "crass" are proposed as subfamily names. These include: *Crudevirinae* (from *crude*), *Obtuvirinae* (from *obtuse*), *Churivirinae* (from *churish/churlish*), *Coarsevirinae* (from *coarse*), *Doltivirinae* (from *doltish*), *Loutivirinae* (from *loutish*), *Uncouvirinae* (from *uncouth*), *Bearivirinae* (from *bearish*), *Boorivirinae* (from *boorish*), *Oafivirinae* (from *oafish*).

Creating genus names for the enormous diversity of uncultured genera in the Earth virosphere is a highly challenging task, which we believe can be carried out efficiently only by using algorithmic approaches. Because many crAss-like phage clades seem to have co-evolved with the human host over extended evolutionary time [5,8], for genus-level taxa, we chose to use modified names of dog breeds, the other long-term companions of the human species.

Dog breeds were randomly chosen from publicly available websites, avoiding breeds reliant heavily on place names or physical/descriptive characteristics (e.g. Portuguese Water Dog).

Each letter of the term to be mutated was assigned a random number, 1-3. Characters were then passed through a three step script to create the following changes:

1. '*a1'→'ah', 'b1'→'p', 'c1'→'k', 'd1'→'t', 'e1'→'eh', 'g1'→'j', 'i1'→'ih', 'k1'→'c', 'o1'→'oh', 't1'→'d', 'u1'→'uh'*
2. *'a2'→'e', 'e2'→'i', 'i2'→'o', 'o2'→'u', 'u2'→'a'*
3. *'a3'→'i', 'e3'→'o', 'i3'→'u', 'o3'→'a', 'u3'→'e', 'j3'→'g', 'k3'→'c', 'p3'→'b', 't3'→'d'*

In order to maximize the likelihood that the mutated term was pronounceable, only the first occurrence of a repeating letter was kept. All long terms were cut after the 7th character and shortened further if needed to the last occurring vowel. This was to prevent a hard consonant before the genus level suffix '*-virus*'. Each mutated word needed to be a minimum of 5 characters in length, have two consonants and two vowels. All names were checked against the ICTV 2019v1 Master List to check if the final generated term matches any currently used taxon name (Realm *→* Species).

**Historical aspects:**

The great majority of the viruses included in the proposed order *Crassvirales* are uncultured, with a few exceptions [9,10,11], and were discovered through assembly of metagenomic sequencing reads into contigs with subsequent automated annotation. The first representative to be discovered and reported in the literature was crAssphage, a ubiquitous and highly prevalent member of the human gut virome [2]. The prototypical crAssphage is detectable in up to 73% of the human faecal metagenomes and can account for up to 90% of virus-like particle enriched metagenomes or 22% reads of the total metagenomes in extreme cases [2]. With other members of the proposed order included, *Crassvirales* are detectable in 98% of faecal viromes from Western cohorts and in 77% of the human faecal viromes worldwide. In 8% of cases crAss-like phages accounted for >50% of the virome, making this group the most abundant viral taxon of the human virome [7].

Biological characteristics of these highly abundant viruses remain, to a large extent, a subject of speculation because the vast majority of species have not been isolated in culture. It has been predicted from the analysis of the genomes of the two prototypical species (crAssphage and ﻿immunodeficiency-associated stool (IAS) virus, [12]) that these viruses likely have *Podoviridae*-like morphology, are virulent and infect bacterial hosts in the phylum Bacteroidetes [3,4]. Shortly afterwards, the first cultured representative ΦCrAss001 was isolated that is related with IAS virus, infecting *Bacteroides intestinalis* [9]. Electron micrographs confirm a podovirus morphology (**Fig. 1**). This phage, although not lysogenic, demonstrates a benign form of interaction with its host, which allows for the continuous proliferation of both host and phage in high numbers. Future research is expected to uncover the molecular details of this unusual interaction and establish whether it underpins the ability of crAss-like phages for stable, long-term and high-level colonization of the mammalian gut.

A close up of a person's eyes

Description automatically generated with medium confidence

***Figure 1.*** *TEM image of uranyl acetate negatively contrasted ΦcrAss001 virions, scale bar is 100 nm (×62,000 magnification, accelerating voltage of 120 kV).*

Given their unique genome architecture, the partial conservation of gene order and monophyly of the most conserved protein-coding genes, recent publications have suggested that crAss-like phages should be unified under a novel, order-level taxon with two or more families included [5]. Here we propose the creation of the novel order *Crassvirales*, comprising four new families, ten new subfamilies, 42 new genera and a total of 73 new species.

**Phylogenomics and genome-based taxonomy of *Crassvirales*:**

In order to explore the diversity and phylogeny of uncultured members of the *Crassvirales* and to infer the genome-based taxonomy of the new viral order, we mined publicly available databases for new crAss-like phage genomes. Our search included NCBI GenBank Nucleotide and WGS repositories, Joint Genome Institute IMG/VR viral contig database and the Gut Virome Database [13,14]. We also re-assembled datasets of NGS reads from a number of published and unpublished human-, animal-host and environmental virome studies, available through NCBI SRA database (as described before [7]).

The resulting pool of contigs was screened for crAss-like phages using the following criteria: (i) a minimum of 50kb in length, (ii) BLASTn identity of >70% and overall alignment length of >3kb (with E-value cut-off of 1e-10) to previously well-defined crAss-like phages [6,7] (iii) no ambiguous nucleotides (i.e. no 'N' bases), (iv) gene coding density >80% using either of the translational code tables 11 or 15. The final database was made non-redundant at 99.9% nucleotide identity and included 1,576 crAss-like phage genomes (partial and complete). To increase robustness of our classification, we further narrowed down the database to include only complete circularly permuted genomes (n=537).

Orthologous gene sharing was analyzed for all identified crass-like phages. Open reading frames (ORFs) were identified using Prodigal v2.6.3 (**Fig. 2**). Amino acid sequences of ORF protein products were then clustered into orthologous groups using OrthoMCL (MCL-clustering of pairwise BLASTp-based distances, [15]). Genomes were clustered hierarchically using Ward.D2 (implemented in R v3.6.6 base package) algorithm with Euclidean distances, based on the orthologous gene presence/absence matrix. PAM-clustering (R package *cluster* v2.1.0) produced a nearly identical result.

Amino acid sequences of four proteins (large terminase subunit, portal, major capsid protein, DNA primase) highly conserved across the order were aligned using MUSCLE v3.8.31 and manually curated before phylogeny reconstruction: truncated sequences (partial CDS at contig ends, intron insertion etc.) were removed, numerous inteins were removed from alignments by identifying and removing the columns present in <10% of individual sequences. The large terminase (TerL) and major capsid protein (MCP) phylogenies were inferred using PhyML v3.3.3 with JTT substitution model (VT model selected as optimal by IQtree was also tested and gave nearly identical result) and approximate likelihood branch support.

Table

Description automatically generated

***Figure 2.*** *Examples of Crassvirales genomes belonging to four different families, randomly picked from the database of 537 complete genomes (non-permuted). ORFs were predicted using Prodigal v2.6.3 and conserved proteins were identified using BLASTp (E-value <1e-10) against a database of protein products of ten representative crAss-like phages functionally annotated using HHpred suite (https://toolkit.tuebingen.mpg.de/tools/hhpred) as described in [7].*

***Families***

The four families (*Intestiviridae*, *Crevaviridae*, *Suoliviridae*, *Steigviridae*) correspond to the four main branches seen in the ML-trees of MCP and TerL proteins (**Fig. 3**), and also four main clusters of gene sharing, each of which comprises a set of genomes sharing at least 17% of orthologs in all possible pairwise comparisons (**Fig. 4**).

Diagram, schematic

Description automatically generated

***Figure 3.*** *Maximum likelyhood phylogenetic tree of 515 Crassvirales MCP and TerL amino acid sequences (after manual curation and removal of numerous inteins). Branch support values are approximate likelyhoods. Tips are coloured by newly assigned sub-families (or families, colours are consistent with Fig. 4 and 5). Main first order branches are labelled with family names. MCP and TerL sequences encoded by a distantly related Cellulophaga phage phi13:2 (NC\_021803) were used as outgroups.*

***Subfamilies***

The subfamily-level taxa were created based on clear monophyletic groups of sister genera clades in the MCP and TerL tree (**Fig. 3,** with a few exceptions in TerL, which have likely resulted from recent horizontal transfer of *terL* genes) and further supported by gene sharing data. Subfamily assignment corresponded very well with hierarchical clustering of genomes based on gene sharing: a minimum of 27-79% of proteins shared within each subfamily in all possible pairwise comparisons (**Fig. 4**). A total of 10 subfamilies were proposed which roughly correspond with formerly proposed candidate genera (I-X), with exception of former candidate genus VI which was reformed into a single family *Steigviridae* with no clearly definable subfamilies in it) [7]. Three out of four families, *Intestiviridae*, *Crevaviridae*, *Suoliviridae*, contain between 2 and 5 subfamilies. Specifically, *Intestiviridae* separates into *Crudevirinae*, *Obtuvirinae* and *Churivirinae, Crevaviridae* separates into *Coarsevirinae* and *Doltivirinae*, while *Suoliviridae* contains *Loutivirinae, Uncouvirinae, Bearivirinae, Boorivirinae* and *Oafivirinae* (**Fig. 3 and 4**).

Chart, treemap chart

Description automatically generated

***Figure 4.*** *Hierarchical clustering of 537 complete Crassvirales genomes, based on the percentage of shared orthologous protein-coding genes (Euclidean distances, Ward.D2 clustering algorithm) and showing separation into families (≥17% shared proteins), subfamilies (27-79% shared proteins) and genera (≥80% shared proteins).*

***Genera***

The genus-level taxa (n=42) were defined from the branching of the MCP and TerL ML-trees (**Fig. 3 and 5**) which is compatible with the topologies of the primase and portal trees (not shown). This division closely matches the grouping established with a flat 80% cut-off of shared orthologous genes (**Fig. 4**). The prototypical crAssphage is included in the genus *Carjivirus* of subfamily *Crudevirinae*, family *Intestiviridae*, which is the single most abundant viral genus in our dataset (133 out of 537 complete genomes belong to this genus). IAS virus is placed into genus *Paundivirus* of family *Steigviridae*, while the first cultured mammalian crAss-like phage ΦCrAss001, infecting *B. intestinalis*, represents genus *Kehishuvirus* of the same family. Bacteriophages DAC15 and DAC17 infecting *Bacteroides thetaiotaomicron* [10] are classified into genus *Wulfhauvirus* of family *Steigviridae*. A recently isolated phage crAss002 infecting *Bacteroides xylanisolvens* [11] belongs to genus *Jahgtovirus* of subfamily *Churivirinae* and family *Intestiviridae*. At least one genome, *Azobacteroides* phage ProJPt-Bp1 (AP017903) forms a separate clade, which cannot be reliably placed into our four-family taxonomy of *Crassvirales*, and will be reserved for future taxonomy updates.

Diagram, schematic

Description automatically generated

***Figure 5.*** *Maximum likelyhood phylogenetic tree of Crassvirales MCP and TerL amino acid sequences encoded by 73 species identified in this study. Branch support values are approximate likelyhoods. Tips are coloured by newly assigned sub-families (or families, colours are consistent with Fig. 3 and 4). Main first order branches are labelled with family names. MCP and TerL sequences encoded by a distantly related Cellulophaga phage phi13:2 (NC\_021803) were used as outgroups.*

***Species***

Species-level demarcation was performed using a flat cut-off of 95% ANI over 85% of contig length. A total of 73 species were identified, having between 1 and 128 representative genomes (128 in *Carjivirus communis* – the prototypical crAssphage), of the 537 complete *Crassvirales* genomes identified in this study. A detailed list of genera and species is provided in the accompanying Excel module. Type strains within each species were selected either because they have been cultured (crAss001, crAss002, DAC15 etc.) or where a substantial amount of previous amount of work has been performed (crAssphage, IAS virus), or (for species with no cultured or well-studied representatives) on the basis of having a genome size closest for the median genome size across representatives of a given species. Species are given binomial species names consisting of the genus name with a freeform species epithet.

**References**

[1] Roux S, Adriaenssens EM, Dutilh BE et al (2019) Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol 37:29-37. doi: 10.1038/nbt.4306. PMID: 30556814

[2] Dutilh BE, Cassman N, McNair K et al (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498. doi: 10.1038/ncomms5498. PMID: 25058116

[3] Dutilh BE, Schmieder R, Nulton J et al (2012) Reference-independent comparative metagenomics using cross-assembly: crAss, Bioinformatics 28: 3225-3231. doi: 10.1093/bioinformatics/bts613. PMID: 23074261

[4] Reyes A, Haynes M, Hanson N et al (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338. doi: 10.1038/nature09199. PMID: 20631792

[5] Koonin EV, Yutin N (2020) The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. Trends Microbiol 28:349-359. doi: 10.1016/j.tim.2020.01.010. PMID: 32298613

[6] Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV (2018) Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 3:38-46. doi: 10.1038/s41564-017-0053-y. PMID: 29133882

[7] Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, Draper LA, Gonzalez-Tortuero E, Ross RP, Hill C (2018) Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host Microbe 24:653-664.e6. doi: 10.1016/j.chom.2018.10.002. PMID: 30449316

[8] Edwards RA, Vega AA, Norman HM et al (2019) Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol 4:1727-1736. doi: 10.1038/s41564-019-0494-6. PMID: 31285584

[9] Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C (2018) ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun 9:4781. doi: 10.1038/s41467-018-07225-7. PMID: 30429469

[10] Hryckowian AJ, Merrill BD, Porter NT, Van Treuren W, Nelson EJ, Garlena RA, Russell DA, Martens EC, Sonnenburg JL (2020) Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. Cell Host Microbe 28:371-379.e5. doi: 10.1016/j.chom.2020.06.011. PMID: 32652063

[11] Guerin E, Shkoporov AN, Stockdale SR, Colom Comas J, Khokhlova EV, Clooney AG, Daly KM, Draper LA, Stephens N, Scholz D, Ross RP & Hill C (2021) Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. BMC Microbiome 9: 89. doi: 10.1186/s40168-021-01036-7. PMID: 33845877.

[12] Oude Munnink BB, Canuti M, Deijs M et al (2014) Unexplained diarrhoea in HIV-1 infected individuals. BMC Infect Dis 14:22. doi: 10.1186/1471-2334-14-22. PMID: 24410947

[13] Gregory AC, Zablocki O, Howell A, Bolduc B, Sullivan MB (2019) The human gut virome database. Preprint from bioRxiv, 31 May 2019. DOI: 10.1101/655910 PPR: PPR80789

[14] Paez-Espino D, Roux S, Chen IA et al (2019) IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes Nucleic Acids Res 47:D678–D686. https://doi.org/10.1093/nar/gky1127 PMID: 30407573

[15] Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178-2189. DOI: 10.1101/gr.1224503 PMID: 12952885