

The International Committee on Taxonomy of Viruses

Taxonomy Proposal Form, 2024

**Part 1a: Details of taxonomy proposals**

|  |  |  |
| --- | --- | --- |
| **Title:** | Promoting the family *Autographiviridae* to create one new order, *Autographivirales*, with four new families, four new subfamilies, 93 new genera and 607 new species (*Duplodnaviria, Caudoviricetes*). | |
| **Code assigned:** | 2024.045B |

|  |  |  |  |
| --- | --- | --- | --- |
| **Author(s), affiliation and email address(es):** | | | |
| **Name** | **Affiliation** | **Email address** | **Corresponding author(s)** X |
| **Turner D** | **School of Applied Sciences, University of the West of England, UK** | **dann2.Turner@uwe.ac.uk** | **X** |
| **Carrillo D** |  | **dcarrillo.bioinf@gmail.com** |  |
| **Lood C** | **University of Oxford, UK** | **cedric.lood@biology.ox.ac.uk** |  |
| **Ely B** | **University of South Carolina, USA** | **ely@sc.edu** |  |
| **Lehman SM** | **Food and Drug Administration, USA** | **susan.lehman@fda.hhs.gov** |  |
| **Dutilh B** | **Utrecht University, Netherlands; Friedrich Schiller University, Jena, Germany** | **bedutilh@gmail.com** |  |
| **Kropinski AM** | **University of Guelph, Canada** | **phage.canada@gmail.com** |  |
| **Lavigne R** | **KU Leuven, Belgium** | **Rob.lavigne@kuleuven.be** |  |
| **Adriaenssens EM** | **Quadram Institute Bioscience, UK** | **Evelien.adriaenssens@quadram.ac.uk** |  |
| **Millard AD** | **University of Leicester** | **Adm39@leicester.ac.uk** |  |

**Part 1b: Taxonomy Proposal Submission**

|  |  |  |  |
| --- | --- | --- | --- |
| **ICTV Subcommittee:** | | | |
| Animal DNA Viruses and Retroviruses |  | Bacterial viruses | **X** |
| Animal minus-strand and dsRNA viruses |  | Fungal and protist viruses |  |
| Animal positive-strand RNA viruses |  | Plant viruses |  |
| Archaeal viruses |  | General - |  |

|  |
| --- |
| **List the ICTV Study Group(s) that have seen or have been involved in creating this proposal:** |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Optional – complete only if formally voted on by an ICTV Study Group:** | | | |
| **Study Group** | **Number of members** | | |
| **Votes in support** | **Votes against** | **No vote** |
|  |  |  |  |
|  |  |  |  |

|  |  |
| --- | --- |
| **Submission date:** | 21/06/2024 |

**Part 1c: Feedback from ICTV Executive Committee (EC) meeting**

|  |  |
| --- | --- |
| **Executive Committee Meeting Decision code:** | **X** |
| A – Accept |  |
| Ac – Accept subject to revision by relevant subcommittee chair. No further vote required | **x** |
| U – Accept without revision but with re-evaluation and email vote by the EC |  |
| Uc – Accept subject to revision and re-evaluation and email vote by the EC |  |
| Ud – Deferred to the next EC meeting, with an invitation to revise based on EC comments |  |
| J - Reject |  |
| W - Withdrawn |  |

|  |
| --- |
| **Comments from the Executive Committee:** |
| Remove the capital letter from where species epithets are words. |

**Part 1d: Revised Taxonomy Proposal Submission**

|  |
| --- |
| **Response of proposer:** |
| Corrected |

|  |  |
| --- | --- |
| **Revision date:** | 30/09/2024 |

**Part 2:** **GENERAL PROPOSAL**

|  |
| --- |
| **Abstract for General Proposal:** |
| *Brief description of current situation:*  *Proposed changes:*  *Justification:* |

|  |
| --- |
| **Text of General Proposal:** |
| *Background:*  *Proposed* *changes:*  *Justification:* |

|  |
| --- |
| **References:** |
|  |

|  |
| --- |
| **Tables, Figures:** |

<Start here>

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

|  |
| --- |
| **Name of accompanying Excel module:** |
| 2024.045B.A.v2.Autographivirales.xlsx |

|  |  |  |  |
| --- | --- | --- | --- |
| **Taxonomic changes proposed:** | | | |
| Establish new taxon | **X** | Split taxon |  |
| Abolish taxon | **X** | Merge taxon |  |
| Move taxon | **X** | Promote taxon | **X** |
| Rename taxon | **X** | Demote taxon |  |
| Move and rename | **X** |

|  |  |  |
| --- | --- | --- |
| **Is any taxon name used here derived from that of a living person:** | | **N** |
| **Taxon name** | **Person from whom the name is derived** | **Attached X** |
|  |  |  |
|  |  |  |
|  |  |  |

|  |
| --- |
| **Abstract of Taxonomy Proposal:** |
| *Taxonomic rank(s) affected*:  Realm: *Duplodnaviria*; Kingdom: *Heunggongvirae*; Phylum: *Uroviricota*; Class: *Caudoviricetes*  *Description of current taxonomy*:  The family *Autographiviridae* was established under taxonomic proposal [2019.103B](https://ictv.global/ictv/proposals/2019.103B.zip)  *Proposed* *taxonomic change(s)*:  We propose;   1. The establishment of a new order, *Autographivirales*, containing four new families. 2. The creation of four new subfamilies 3. The creation of 93 new genera 4. The creation of 610 new species 5. Abolition of 21 species   *Justification*:  The proposed order forms a single deep-branching clade in tBLASTx distance analysis, reflected in core gene maximum-likelihood phylogeny. The proposed families form monophyletic clusters in proteome-based analyses and each share a number of core orthologous genes. |

|  |
| --- |
| * **Text of Taxonomy proposal:** |
| *Taxonomic rank(s) affected*:  Realm: *Duplodnaviria*; Kingdom: *Heunggongvirae*; Phylum: *Uroviricota*; Class: *Caudoviricetes*  *Description of current taxonomy*:  The *Autographiviridae* family was established under taxonomic proposal 2019.103B.  *Proposed* *taxonomic change(s)*:  We propose the establishment of a new order, *Autographivirales*, containing four new families, four new subfamilies, 93 new genera and 607 new species.  *Demarcation criteria:*  **Genus and species** were demarcated from pairwise intergenomic distances calculated using the similarity function of tax\_myPhage (<https://github.com/amillard/tax_myPHAGE>) and employing current ICTV BVS demarcation criteria where species are defined as genomes exhibiting ≤95% similarity and genera as possessing ≥70% similarity over their genome length [12].  **Subfamilies** were demarcated where bacterial virus genomes exhibited low nucleotide sequence similarity (~35-50%), shared a high number of protein clusters, and formed clades in signature gene phylogenies.  **Families** were demarcated by the presence of (i) a deep branching clade in hierarchical tblastx distance analysis, (ii) the presence of shared core genes ranging from four to twelve depending upon the family, and (iii) formation of monophyletic clades in partitioned maximum-likelihood phylogenetic trees.  **Orders** were demarcated by searches conducted using HMM profiles identified from MCL clustering of profile-profile comparisons of proteins grouped by MMSeqs2.  *Justification*:  In recent years the number of genome sequences deposited to the INSDC that exhibit similarity to the family *Autographividae* has substantially increased, requiring a reassessment of the current taxonomy in light of these new isolates.  **Dataset generation**  To identify additional phage isolates that belong to the family *Autographviridae* we first examined the Virus Metadata Resource [version MSL38\_v3; <https://ictv.global/vmr>]. Of the 373 species classified, a total of 28 sequence records representing metagenome assemblies, partial genomes and records flagged as “UNVERIFIED” in GenBank [1-2] were excluded, yielding an initial dataset of 348 genomes. To identify conserved “signature” genes across the 348 genomes, protein sequences were clustered using a custom snakemake pipeline. Briefly, proteins were first clustered using MMSeqs2 (release 15-6f452; [3]) and multiple sequence alignments constructed with MAFFT (version 7.520; [4, 5]). Pairwise profile-to-profile comparisons of protein clusters were performed with HHblits (version 3.3.0 [6]) to construct a network of similarities between each group of proteins before MCL (version 14.137; [7]) clustering was used to produce protein super-clusters. We selected 12 protein super-clusters as candidates for “signature genes” based on the presence of a representative protein sequence in >94% of genomes in the initial dataset. HMM profiles corresponding to the 12most highly conserved signature genes were then used to search the INPHARED 3Apr2024 dataset [8]. We expanded the initial dataset to include all genomes that exhibited ≥50% coverage for at least 9 or more of the 12 HMM profiles using hmmsearch (version 3.4; 10.1371 [9]) resulting in a final dataset of 1,468 genomes. This dataset was compared to ICTV classified phage families and a number of proposed families in the class *Caudoviricetes*. As a secondary measure, the beta version (b38) of the viral classification tool vConTACT3 was also used to assess taxa.  **Order**  We propose the creation of a new order, *Autographivirales*. Hierarchical clustering of the predicted proteomes using the ViPTreeGen (version 1.2.2 [10-11]) demonstrates that these genomes form a deep-branching clade (Figure 1, Figure 2). A total of 11 genes were common in over 99% of the genomes, which increased to 100% upon manual inspection to identify genes wrapped around the genome ends. We propose these HMMs represent orthologs that can be used to delineate the proposed order.  **Families**  We propose the creation of four new families, named *Autoscriptoviridae*, *Autonotataviridae, Autotranscriptaviridae* and *Autosignataviridae*. The proposed families form monophyletic clades in single and core proteome maximum likelihood phylogenies (Figure 3). Each family forms a deep-branching clade in hierarchical trees produced using tBLASTx distances in the stand-alone tool ViPTreeGen v1.2.1 (Figure 2, Figure 2). Each family shares several core genes (Tables 4-7).  **Subfamilies**  We propose the creation of four new subfamilies. The data suggest that a number of additional nascent subfamilies exist. However, we have chosen not to create subfamilies where the number of representative genomes numbered less than ten. Single gene phylogeny of the small Terminase subunit results in paraphyly for the subfamilies *Molineuxvirinae* and *Colwellvirinae*, and dispersion of member genomes of the families *Autoscriptoviridae* and *Autonotatavirdae* (Figure 6). We recommend that phylogenetic analysis of this single protein is not used for taxonomic assignment. We have chosen to assign subfamilies based upon genomes sharing ≥30% nucleotide sequence similarity and the distance of clades in the core-gene phylogeny. These assignments will be reviewed in the future with the addition of new genomes and coding complete metavirome sequences. The number, core and shared protein clusters for each subfamily are presented in Table 8. We note that the *Studiervirinae* exhibit a low number of protein clusters using the approach described here and may be over-extended. This subfamily will be reassessed in the future.  **Genera and species**  Based on nucleotide sequence similarity we propose the creation of 91 new genera and 651 new species. A full matrix of genome nucleotide similarities is provided as supplementary material. We propose to abolish 21 species on the grounds that they are either (i) not coding complete (ii) represent strains of previously defined species, or (iii) are listed as “unverified” in GenBank.  **Discrepancies in single gene phylogenies**  While the core-genome phylogeny and hierarchical tree of tBLASTx distances showed consistent groupings, some discrepancies were observed in ML trees of single genes. These discrepancies have been marked in the single gene phylogenies.   1. Terminase, small subunit (Figure 6).   The family *Autonotataviridae* is split.  The subfamilies *Molineuxvirinae* and *Colwellvirinae* are paraphyletic.  The subfamily *Gujervirinae* is split across multiple clades.   1. RNaseH (Figure 7).   The subfamily *Krylovirinae* is split by genomes not classified to the subfamily.     1. Major capsid protein (Figure 11).   The clade encompassing the *Gujervirinae* contains a genome (KX660669) which is not classified as a member of the subfamily.  The clade encompassing the *Slopekvirinae* contains a genome (MT259468) which is not classified as a member of the subfamily.   1. DNA polymerase (Figure 14).   The family *Autonotataviridae* and *Autoscriptoviridae* are interspersed.  This discrepancy may arise from the presence of split coding sequences for the DNA polymerase in some genomes. The sequences were not concatenated prior to calculation of the phylogenetic tree.  *Derivation of names:*  **Order:**  *Autographivirales*. from the Ancient Greek αὐτός (autós), meaning "self" and γρᾰ́φειν (gráphein) meaning "to carve" referring to "self-writing” or “self-transcribing” bacterial viruses in the order that encode a large single subunit RNA polymerase; the suffix -virales for order taxa. This name has been retained to provide continuity and ease of reference for the bacteriophage research community.  **Families**  *Autoscriptoviridae*  The family name is derived from the Latin word “scripto” meaning “to write”. The name uses the neuter plural.  *Autonotataviridae*  The family name is derived from the Latin word “notare” meaning “to inscribe”. The name uses the neuter plural.  *Autotranscriptaviridae*  The family name is derived from the Latin word “transcribere” meaning “to copy”. The name uses the neuter plural.  *Autosignataviridae*  The family name is derived from the Latin word “signare” meaning “to sign”. The name uses the neuter plural.  **Subfamilies**  *Gujervirinae*  This subfamily is named after Grete Kellenberger-Gujer (1919-2011). She made fundamental contributions in the development of research programs in the new field of molecular biology and “that prophage repression, induction and chromosome association (integration) involved separate physiologic events, and her demonstration of genetic recombination by DNA breakage and rejoining, rather than by the alternative and then popular replicative mechanism.” She published collaborative articles with Eduard Kellenberger, Jean Weigle, W. Arber, D. Berg, Robert Weisberg and Anna Podhajska. Further information: <https://www.tandfonline.com/doi/full/10.1080/21597081.2016.1173168>  *Dunnvirinae*  This subfamily is named after John J. Dunn, who collaborated closely with William Studier at Brookhaven National Laboratory where he worked to examine how gene expression is controlled during T7 infection and undertook the first genome sequencing of T7 with Willie Crockett. Obituary: <https://www.bnl.gov/newsroom/news.php?a=23348>.  *Stentvirinae*  This subfamily is named after Gunther Stent (1924-2008). He undertook a postdoctoral fellowship at Max Delbruck’s laboratory at Caltech. Amongst his work at the University of Berkley, California, he used radiolabeled bacteriophage DNA to confirm DNA’s double helix structure and wrote the Molecular Biology of Bacterial Viruses in 1963. Obituary: <https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(08)61059-4.pdf>  *Sechaudvirinae*  Janine Séchaud (d. 2017) was a PhD student and Postdoc with George Streisinger where she pioneered techniques for the study of bacteriophages by transmission electron microscopy and studied the structure of the T4 genome. Further information: <https://www.letemps.ch/opinions/femme-derriere-nobel-suisse-chimie-jacques-dubochet>.  **Genera**  The origin of names for each new genus are described in Table 8. |
|  |

|  |
| --- |
| **References:** |
| 1. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, Marchler-Bauer A, Landrum M, Lathrop S, Lu Z, Madden TL, O'Leary N, Phan L, Rangwala SH, Schneider VA, Skripchenko Y, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021 Jan 8;49(D1):D10-D17. doi: 10.1093/nar/gkaa892. 2. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-45. doi: 10.1093/nar/gkv1189.      1. Steinegger M, Söding J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35(11):1026-1028. doi: 10.1038/nbt.3988. 2. Nakamura T, Yamada KD, Tomii K, Katoh K. (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 34(14):2490-2492. doi: 10.1093/bioinformatics/bty121. 3. Katoh K, Standley DM. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772-80. doi: 10.1093/molbev/mst010. 4. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 20(1):473. doi: 10.1186/s12859-019-3019-7. 5. Van Dongen S. (2008). Graph clustering via a discrete uncoupling process, Siam Journal on Matrix Analysis and Applications 30(1), 121-141. 6. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, Stekel DJ, Hobman J, Jones MA, Millard A. (2021) INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. Phage 2(4):214-223. doi: 10.1089/phage.2021.0007. 7. Eddy SR. (2011) Accelerated Profile HMM Searches. PLoS Comput Biol. 7(10):e1002195. doi: 10.1371/journal.pcbi.1002195. 8. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. (2017) ViPTree: The viral proteomic tree server. Bioinformatics. 33(15):2379–80. 9. Rohwer F, Edwards R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. Journal of Bacteriology. 184(16):4529–35 10. Turner D, Kropinski AM, Adriaenssens EM. (2021) A Roadmap for Genome-Based Phage Taxonomy. Viruses. 13(3):506. doi: 10.3390/v13030506. 11. Letunic I, Bork P. (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics.23(1):127–8. 12. Nguyen LT, Schmidt HA, von Haeseler A, and Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Molecular Biology and Evolution, 32:268-274. https://doi.org/10.1093/molbev/msu300 13. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution, 35:518–522. <https://doi.org/10.1093/molbev/msx281> 14. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, and Jermiin JS (2017) ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates, Nature Methods, 14:587–589. https://doi.org/10.1038/nmeth.4285 |

|  |
| --- |
| **Tables, Figures:** |

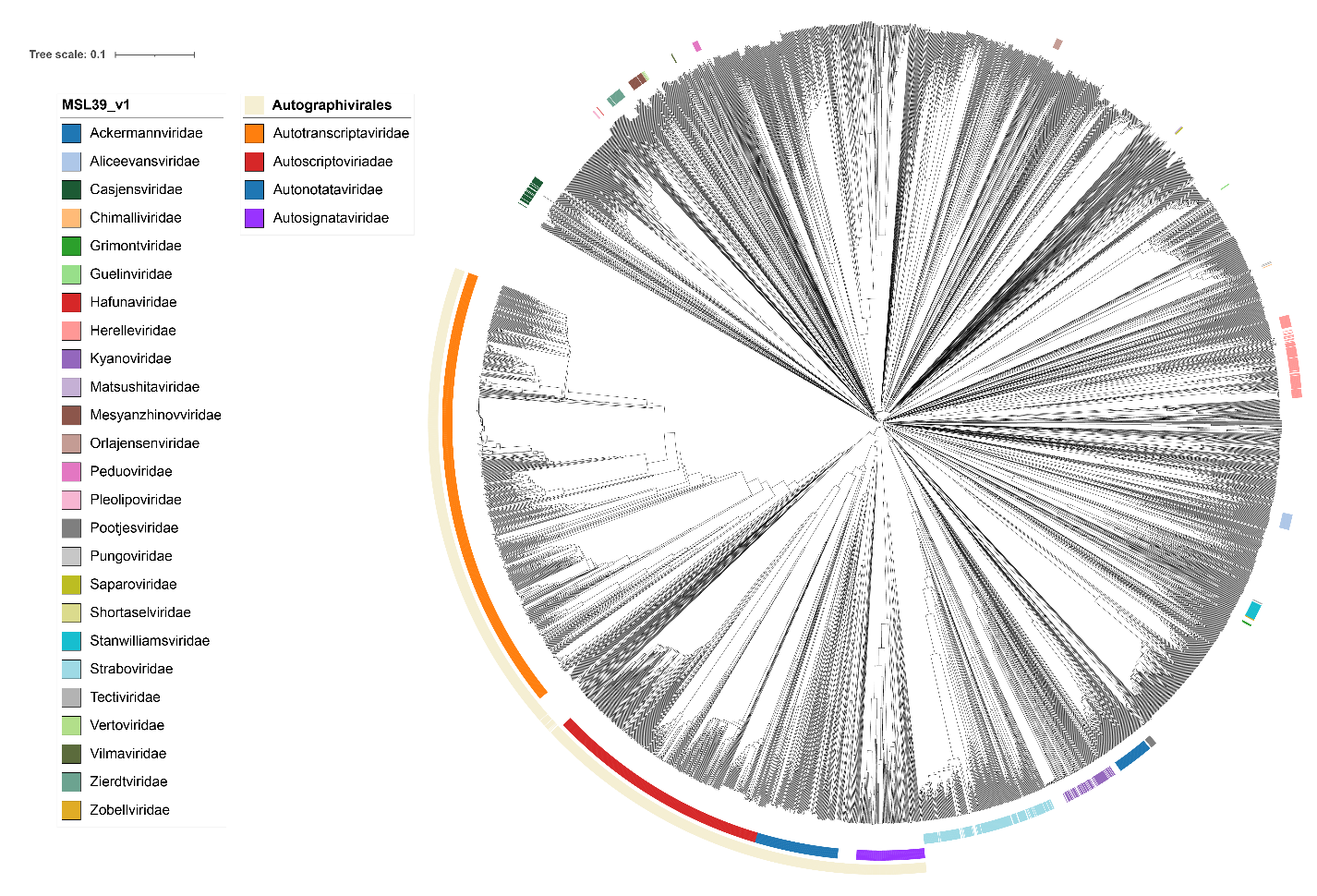
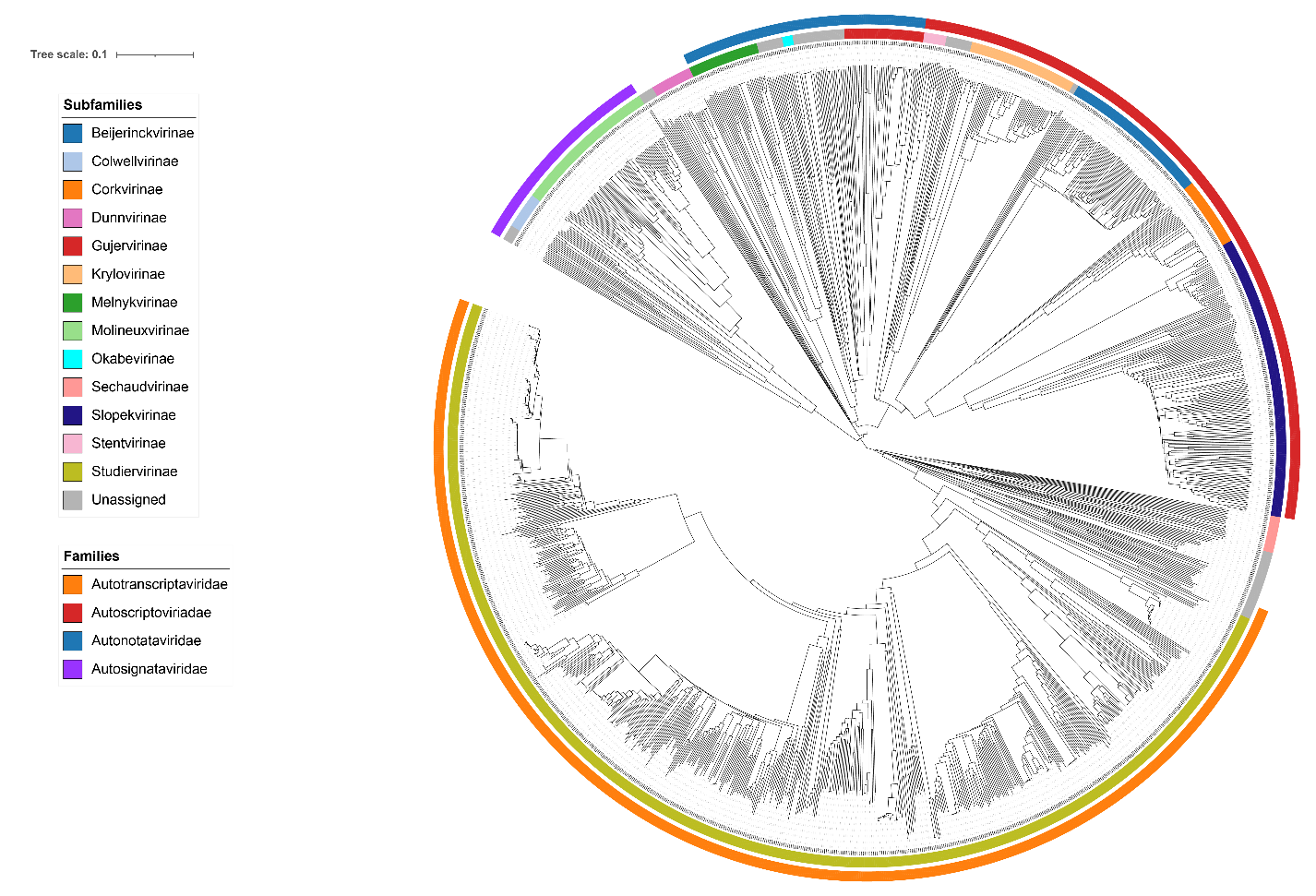


Figure 1. ViPTree proteomic tree of 4,408 bacterial viruses with classified and proposed viral families labeled by the coloured ring. The order *Autographivirales* is demarcated by the outer ring. The proposed families are shown as coloured strips in the middle ring. ICTV classified families of bacterial viruses from MSL39\_v1 are shown as coloured strips in the inner ring. Gaps in these coloured strips arise from the presence of strains, which are not classified by the ICTV. The hierarchical tree was created using ViPTreeGen (version 1.1.2) [10-11], mid-point rooted, and annotated using iToL [14-16]. The tree is based on a dissimilarity matrix generated by pairwise tBLASTx scores between each of the genomes.

Figure 2. ViPTree analysis Proteomic tree of 1,468 bacterial viruses in the proposed order *Autographivirales*. The proposedfamilies and subfamilies are labeled by coloured rings. The hierarchical tree was created using ViPTreeGen (version 1.1.2) [10-11], mid-point rooted, and annotated using iToL [14-16]. The tree is based on a dissimilarity matrix generated by pairwise tBLASTx scores between each of the genomes.

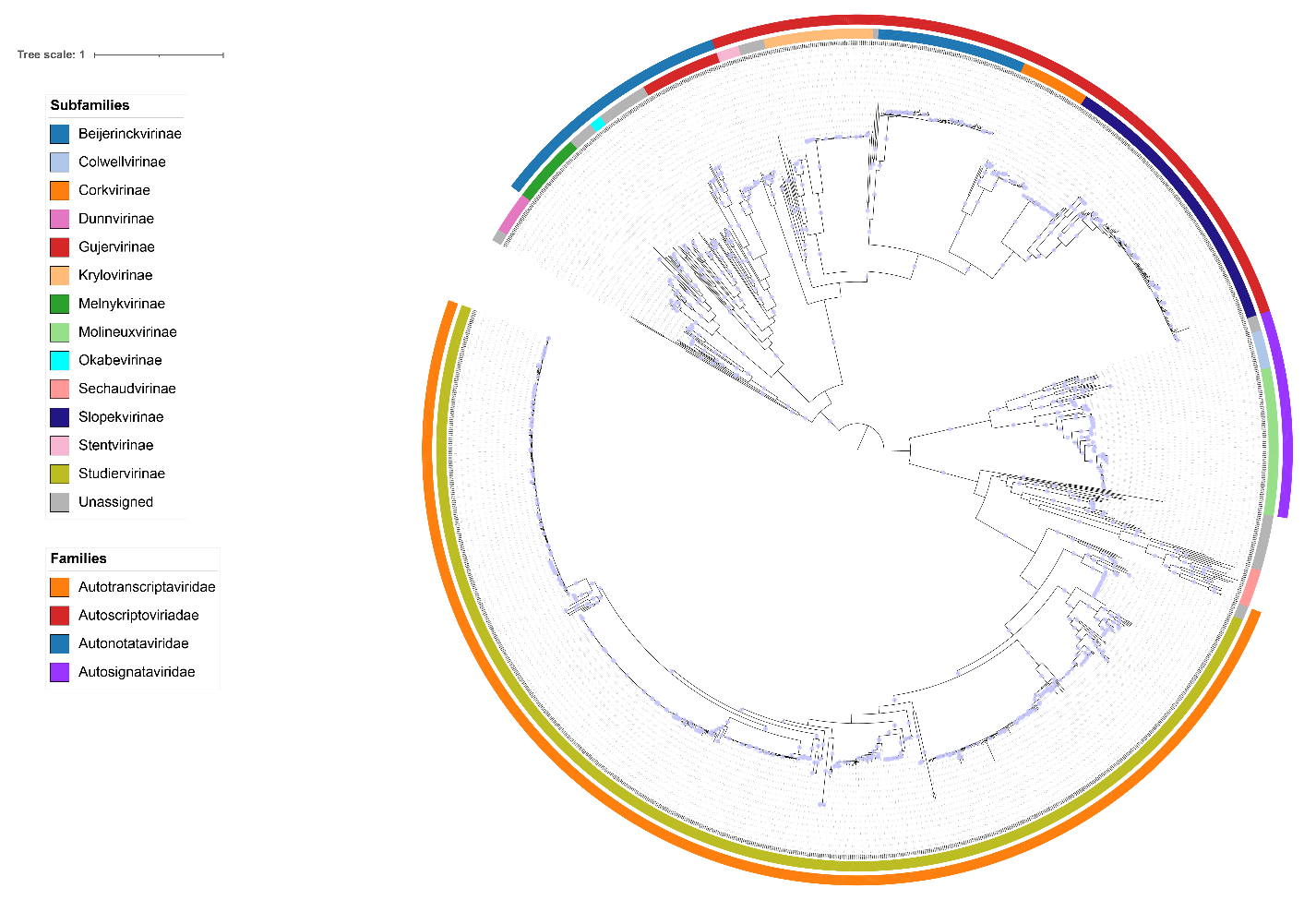
Figure 3. Core-gene phylogeny of the *Autographivirales*. A concatenated protein phylogeny of 11 core genes present in the 1,468 members of the proposed order. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

Table 3. Core genes of the order *Autographivirales*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **HMM profile** | **Number of proteins** | **Number of genomes (1468 total)** | **Percentage of genomes** | **Gene in phage T7 [INSDC accession]** | **Predicted function** |
| 1 | 23451 | 1104 | 72.20% | gene 3 [CAA24402] | endonuclease |
| 2 | 1680\* | 14672 | 99.93% | gene 5 [CAA24412] | DNA polymerase I |
| 3 | 1507\* | 14672 | 99.93% | gene 19 [CAA24440] | Terminase, large subunit |
| 4 | 1503\* | 14662 | 99.86% | gene 8 [CAA24425] | Portal vertex protein |
| 5 | 1492\* | 1468 | 100.00% | gene 7 [CAA24390] | RNA polymerase |
| 6 | 1480\* | 14662 | 99.86% | gene 4A [CAA24405] | DNA primase/helicase |
| 7 | 1474\* | 14672 | 99.93% | gene 10A [CAA24427] | Major capsid protein |
| 8 | 1467 | 14672 | 99.93% | gene 18 [CAA24437] | Terminase, small subunit |
| 9 | 1467 | 14682 | 100.00% | gene 6 [CAA24418] | RNaseH |
| 10 | 1464 | 14602 | 99.46% | gene 12 [CAA24430] | Tail tubular protein B |
| 11 | 1463 | 14642 | 99.73% | gene 11 [CAA24429] | Tail tubular protein A |
| 12 | 1462 | 14632 | 99.66% | gene 9 [CAA24426] | Scaffold |

\*The total number of proteins is greater than the number of phages due to the presence of split coding sequences in a small number of genomes. This protein was not used for the construction of core or single gene phylogenies. 1HMM profile recruits HNH homing endonucleases. 2There is 100% representation of these HMMs when coding sequences split across the genome ends are taken into account.

Table 4. Core genes in *Autotranscriptaviridae*

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein cluster** | **Number of genomes (748 total)** | **Percentage of genomes** | **Predicted product** |
| subfamily\_001 | 748 | 100.00% | RNaseH |
| subfamily\_002 | 747 | 99.87% | DNA polymerase I |
| subfamily\_003 | 747 | 99.87% | Tail protein |
| subfamily\_004 | 747 | 99.87% | Endonuclease |
| subfamily\_005 | 746 | 99.73% | Terminase, small subunit |
| subfamily\_006 | 745 | 99.60% | Terminase, large subunit |
| subfamily\_007 | 745 | 99.60% | RNA polymerase |
| subfamily\_008 | 744 | 99.47% | Portal protein |
| subfamily\_009 | 743 | 99.33% | Tail protein |
| subfamily\_010 | 743 | 99.33% | Amidase |
| subfamily\_011 | 743 | 99.33% | Major head protein |
| subfamily\_012 | 742 | 99.20% | DNA primase/helicase |

Table 5. Core genes in *Autoscriptoviridae*

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein cluster** | **Number of genomes (382 total)** | **Percentage of genomes** | **Predicted product** |
| 1 | 383 | 100.26% | Major head protein |
| 2 | 382 | 100.00% | RNaseH |
| 3 | 380 | 99.48% | Endonuclease VII |
| 4 | 380 | 99.48% | Portal protein |

Table 6. Core genes in *Autonotataviridae*

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein cluster** | **Number of genomes (139 total)** | **Percentage of genomes** | **Predicted product** |
| 1 | 139 | 100.00% | Portal protein |
| 2 | 139 | 100.00% | Terminase, large subunit |
| 3 | 139 | 100.00% | Major head protein |
| 4 | 137 | 98.56% | Tail tubular protein B |

Table 7. Core genes in *Autosignataviridae*

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein cluster** | **Number of genomes (116 total)** | **Percentage of genomes** | **Predicted product** |
| subfamily\_002 | 116 | 100.00% | Hypothetical protein |
| subfamily\_003 | 116 | 100.00% | Tail protein |
| subfamily\_004 | 116 | 100.00% | Terminase, large subunit |
| subfamily\_005 | 116 | 100.00% | DNA polymerase |
| subfamily\_006 | 116 | 100.00% | RNaseH |
| subfamily\_007 | 116 | 100.00% | DNA primase/helicase |
| subfamily\_008 | 116 | 100.00% | ATP-dependent DNA ligase |
| subfamily\_009 | 115 | 99.14% | RNA polymerase |
| subfamily\_010 | 115 | 99.14% | Portal protein |
| subfamily\_011 | 115 | 99.14% | Endonuclease VII |
| subfamily\_012 | 115 | 99.14% | Major head protein |
| subfamily\_013 | 115 | 99.14% | Terminase, small subunit |
| subfamily\_014 | 112 | 96.55% | Hypothetical protein |

Table 8. Summary of protein clusters (PCs) by subfamily. The total number of PCs present in each subfamily and core PCs (present in all members of the subfamily) are shown. The heatmap shows the percentage of PCs shared with other subfamilies relative to the total number of PCs.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | **Percentage of PCs shared between subfamilies** | | | | | | | | | | | | |
| **Subfamily** | **N. of Genomes** | **Total number of PCs** | **Number of Core PCs** | ***Slopekvirinae*** | ***Beijerinckvirinae*** | ***Krylovirinae*** | ***Molineuxvirinae*** | ***Corkvirinae*** | ***Gujervirinae*** | ***Melnykvirinae*** | ***Stentvirinae*** | ***Colwellvirinae*** | ***Okabevirinae*** | ***Dunnvirinae*** | ***Sechaudvirinae*** | ***Studiervirinae*** |
| *Slopekvirinae* | 162 | 309 | 10 | 100.00 | 22.94 | 22.76 | 23.14 | 41.96 | 11.38 | 20.12 | 35.71 | 15.04 | 25.00 | 11.11 | 4.28 | 6.27 |
| *Beijerinckvirinae* | 86 | 109 | 19 | 8.09 | 100.00 | 17.89 | 5.68 | 18.88 | 10.78 | 7.10 | 21.43 | 11.28 | 25.00 | 11.11 | 4.81 | 1.82 |
| *Krylovirinae* | 63 | 123 | 13 | 9.06 | 20.18 | 100.00 | 5.68 | 30.07 | 15.57 | 10.65 | 30.00 | 9.02 | 25.00 | 11.11 | 2.14 | 2.81 |
| *Molineuxvirinae* | 85 | 229 | 16 | 17.15 | 11.93 | 10.57 | 100.00 | 16.08 | 9.58 | 14.20 | 14.29 | 24.81 | 19.64 | 8.73 | 7.49 | 9.90 |
| *Corkvirinae* | 40 | 143 | 16 | 19.42 | 24.77 | 34.96 | 10.04 | 100.00 | 13.17 | 11.83 | 31.43 | 9.02 | 25.00 | 11.11 | 2.67 | 3.47 |
| *Gujervirinae* | 46 | 167 | 15 | 6.15 | 16.51 | 21.14 | 6.99 | 15.38 | 100.00 | 12.43 | 20.00 | 9.77 | 44.64 | 13.49 | 4.28 | 2.48 |
| *Melnykvirinae* | 41 | 169 | 14 | 11.00 | 11.01 | 14.63 | 10.48 | 13.99 | 12.57 | 100.00 | 27.14 | 12.78 | 33.93 | 13.49 | 3.21 | 5.12 |
| *Stentvirinae* | 13 | 70 | 30 | 8.09 | 13.76 | 17.07 | 4.37 | 15.38 | 8.38 | 11.24 | 100.00 | 5.26 | 19.64 | 9.52 | 2.14 | 1.16 |
| *Colwellvirinae* | 22 | 133 | 20 | 6.47 | 13.76 | 9.76 | 14.41 | 8.39 | 7.78 | 10.06 | 10.00 | 100.00 | 16.07 | 8.73 | 5.35 | 3.96 |
| *Okabevirinae* | 7 | 56 | 29 | 4.53 | 12.84 | 11.38 | 4.80 | 9.79 | 14.97 | 11.24 | 15.71 | 6.77 | 100.00 | 11.11 | 3.21 | 1.32 |
| *Dunnvirinae* | 24 | 126 | 19 | 4.53 | 12.84 | 11.38 | 4.80 | 9.79 | 10.18 | 10.06 | 17.14 | 8.27 | 25.00 | 100.00 | 4.28 | 1.65 |
| *Sechaudvirinae* | 22 | 187 | 12 | 2.59 | 8.26 | 3.25 | 6.11 | 3.50 | 4.79 | 3.55 | 5.71 | 7.52 | 10.71 | 6.35 | 100.00 | 3.47 |
| *Studiervirinae* | 740 | 606 | 2 | 12.30 | 10.09 | 13.82 | 26.20 | 14.69 | 8.98 | 18.34 | 10.00 | 18.05 | 14.29 | 7.94 | 11.23 | 100.00 |

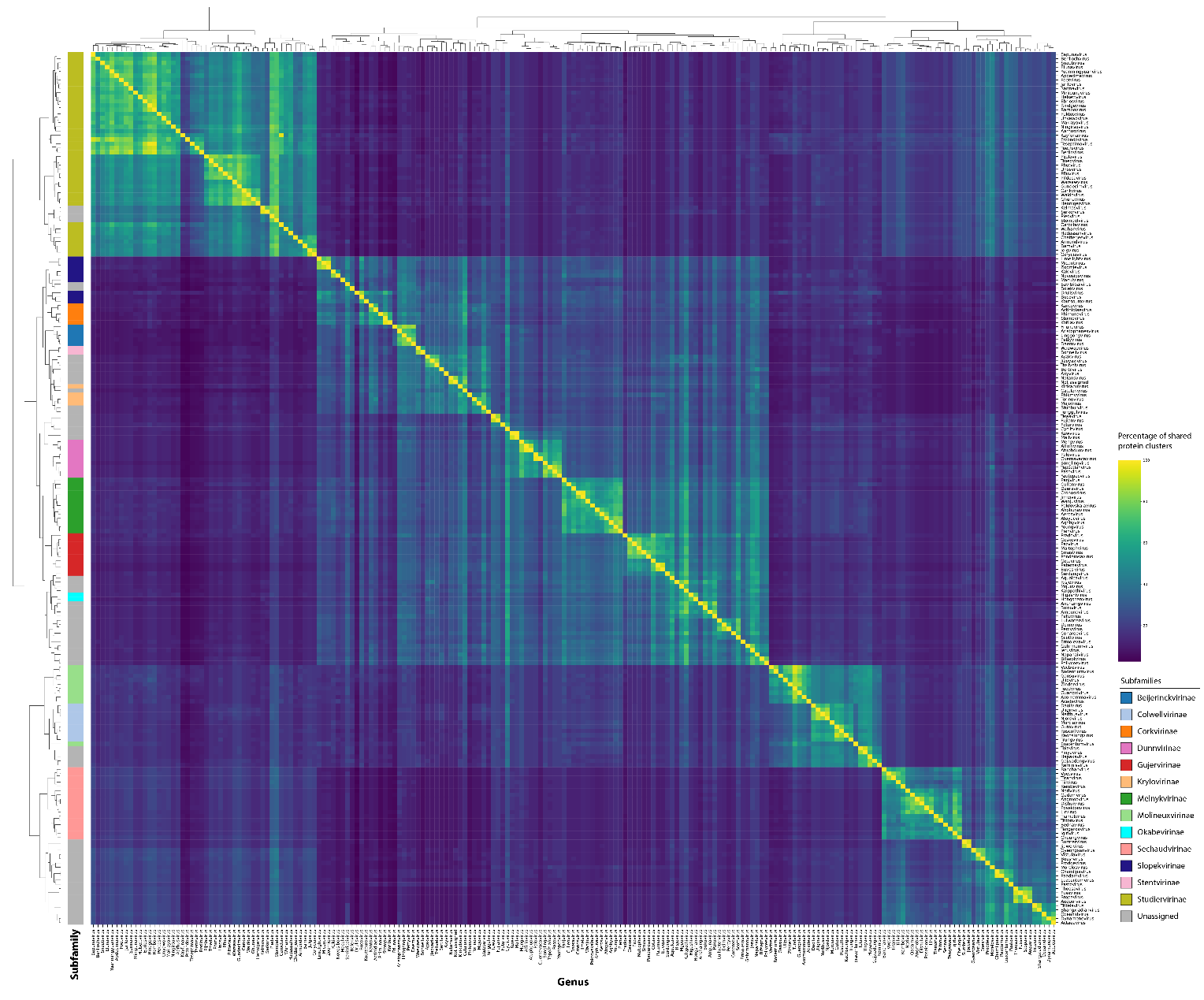


Figure 4. Heatmap of the percentage shared protein clusters between genera of the *Autographivirales*. The heatmap was visualized using Seaborn and hierarchically clustered using the complete linkage method via a custom python script. The coloured bar represents the subfamily each genus is assigned to.

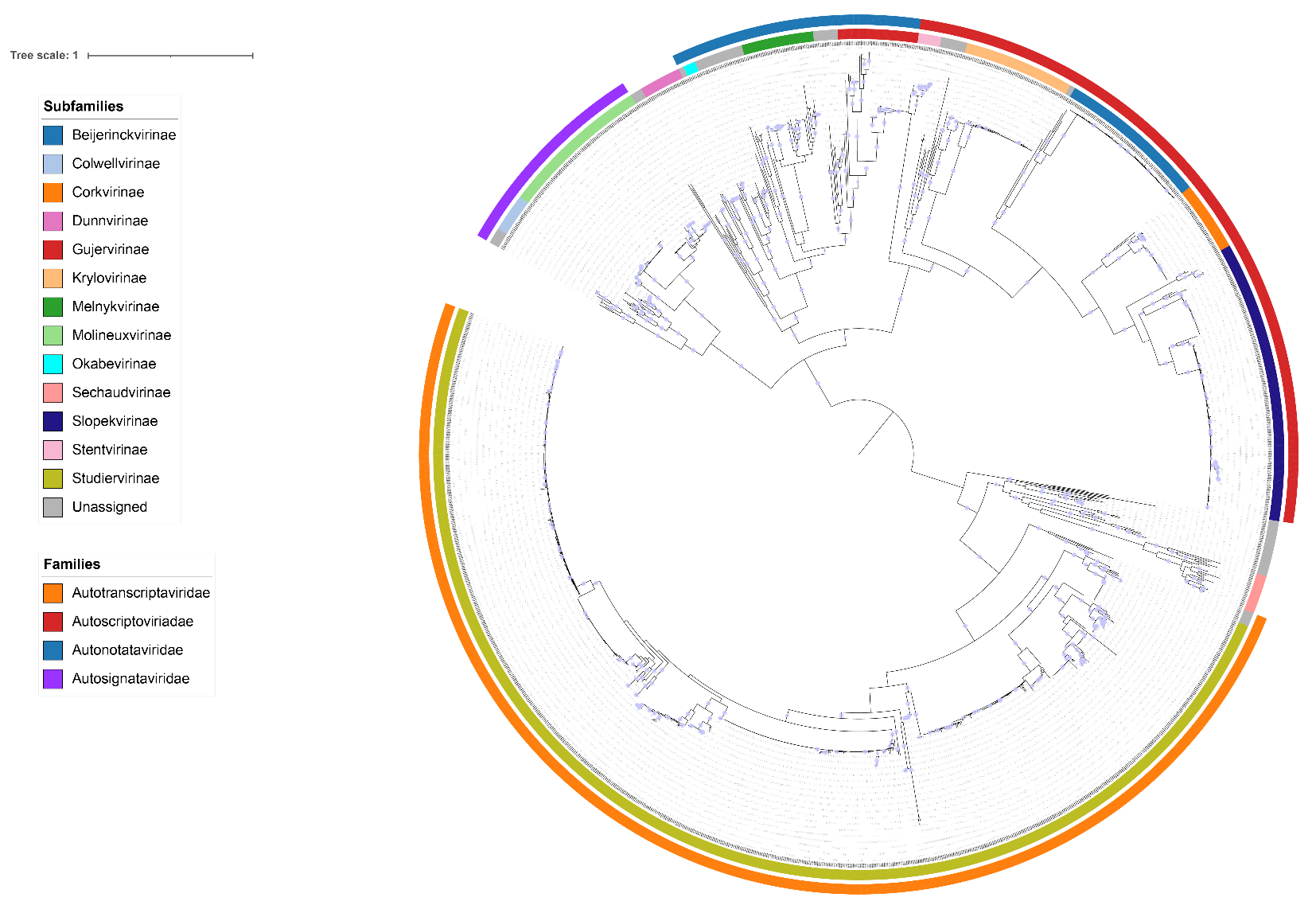


Figure 5. Single gene phylogeny of the large terminase subunit. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

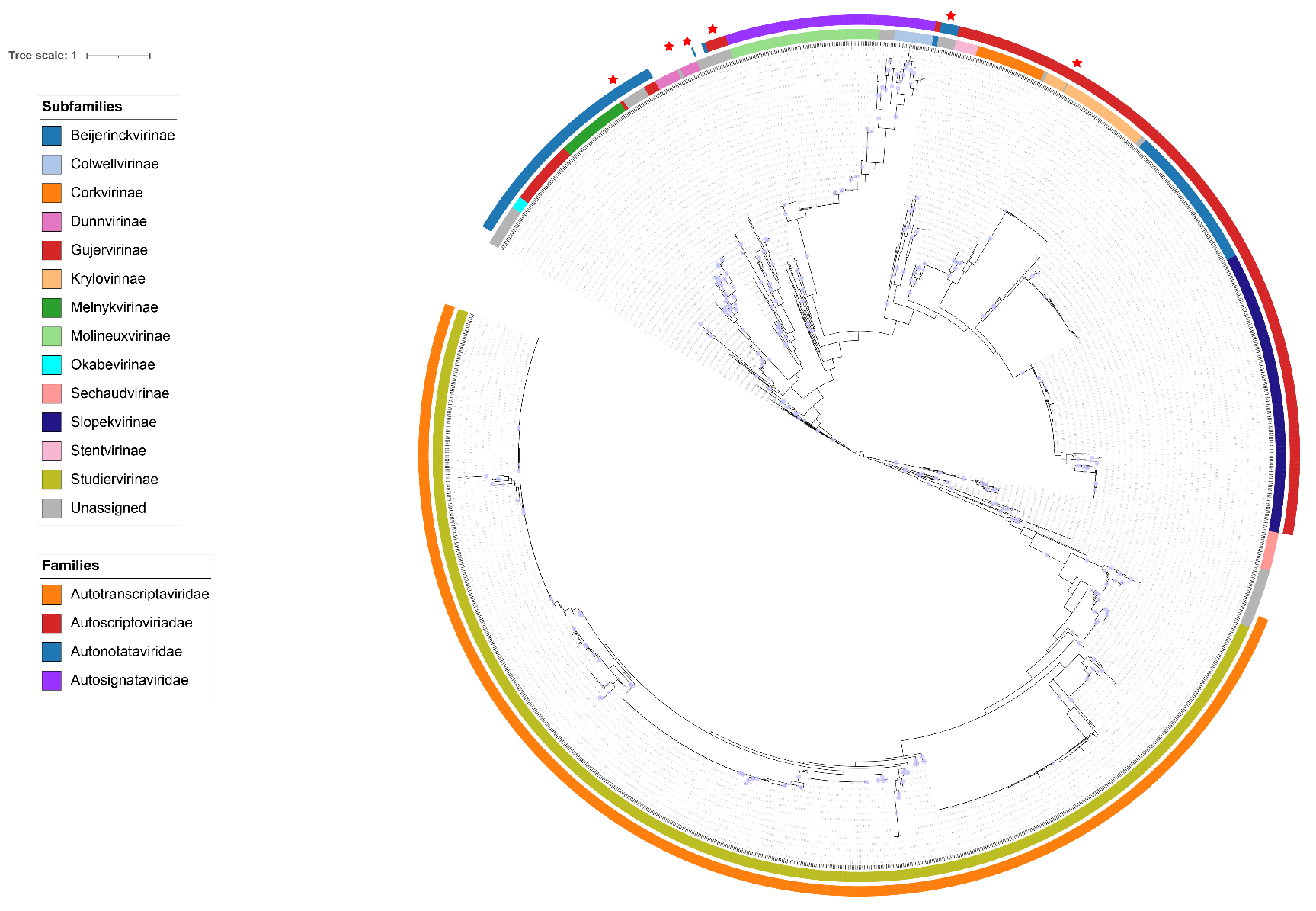


Figure 6. Single gene phylogeny of the small terminase subunit. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers. Discrepancies between the single gene phylogeny and proposed classification are marked by a red star symbol.

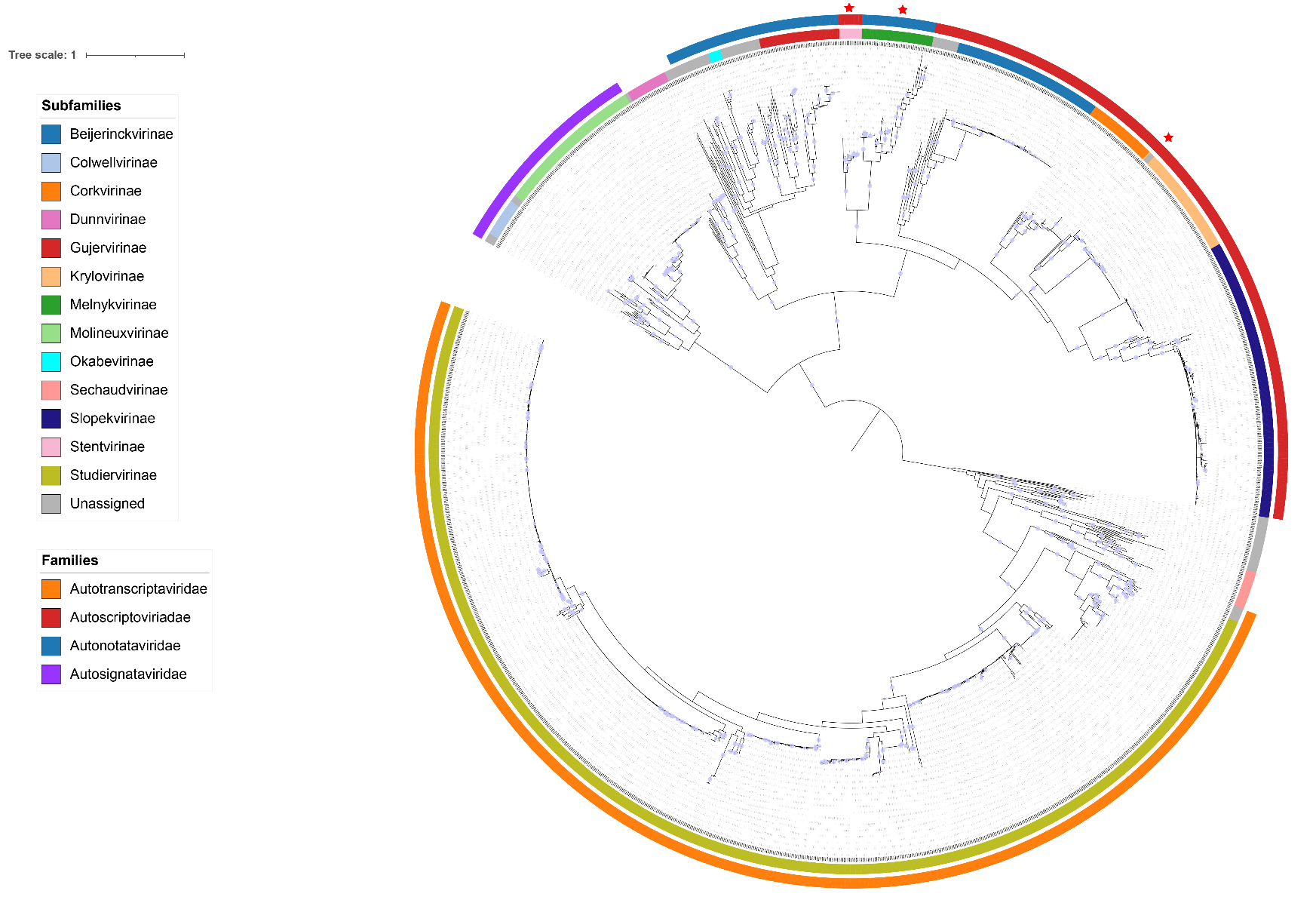


Figure 7. Single gene phylogeny of the RNaseH. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers. Discrepancies between the single gene phylogeny and proposed classification are marked by a red star symbol.

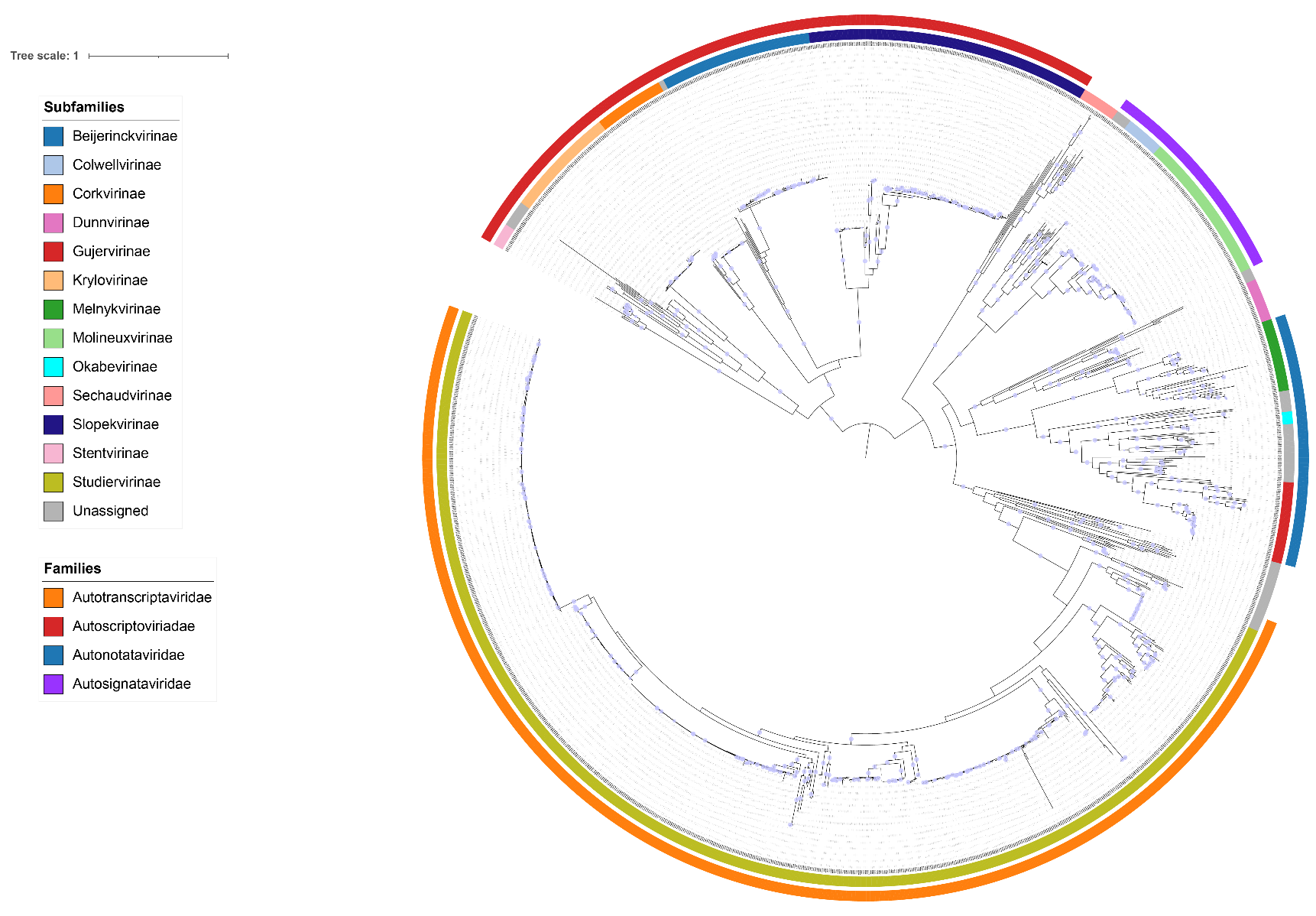


Figure 8. Single gene phylogeny of the RNA polymerase. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

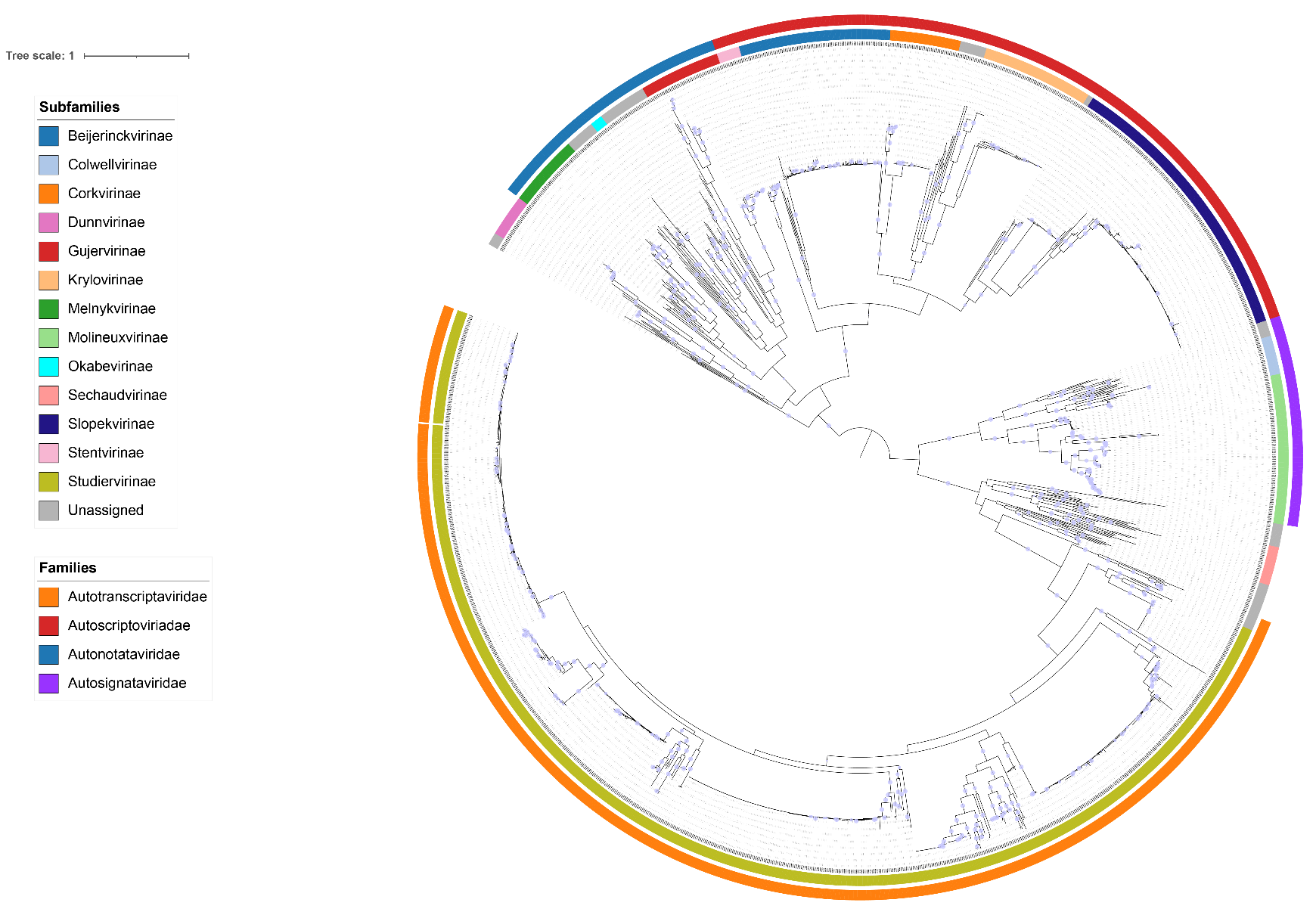


Figure 9. Single gene phylogeny of the capsid scaffold protein. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

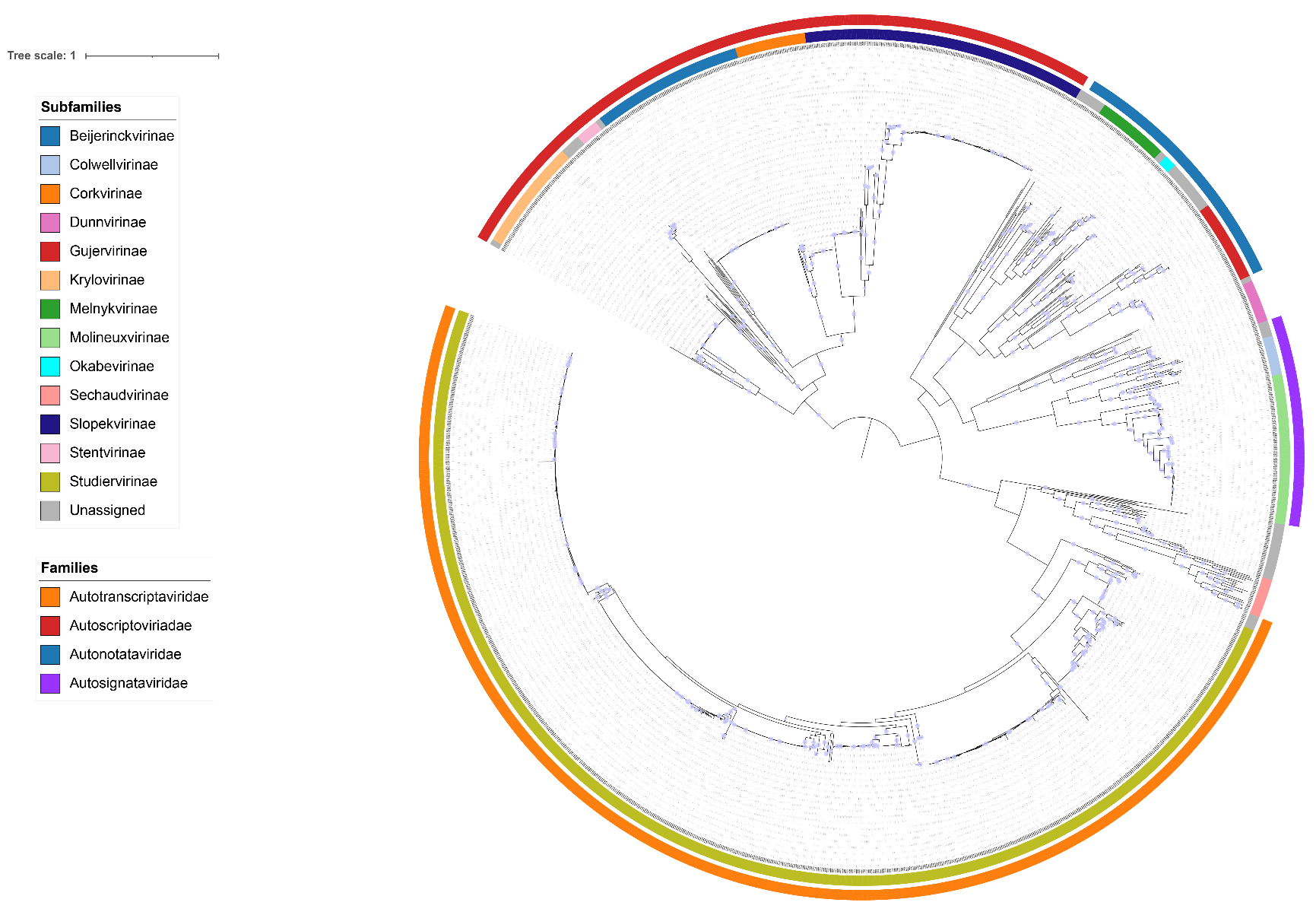
****

Figure 10. Single gene phylogeny of the portal protein. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

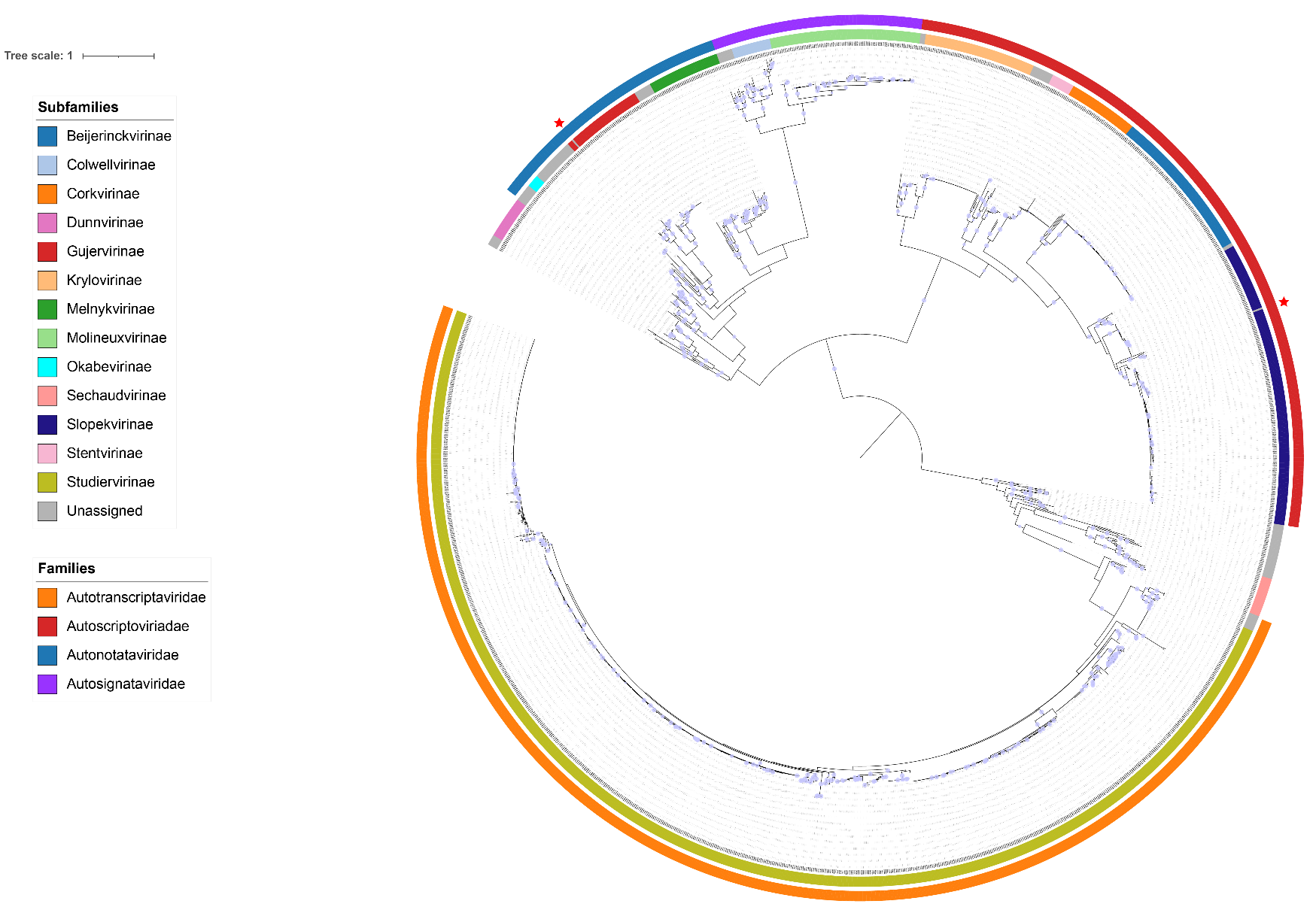
****

Figure 11. Single gene phylogeny of the major capsid protein. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers. Discrepancies between the single gene phylogeny and proposed classification are marked by a red star symbol.

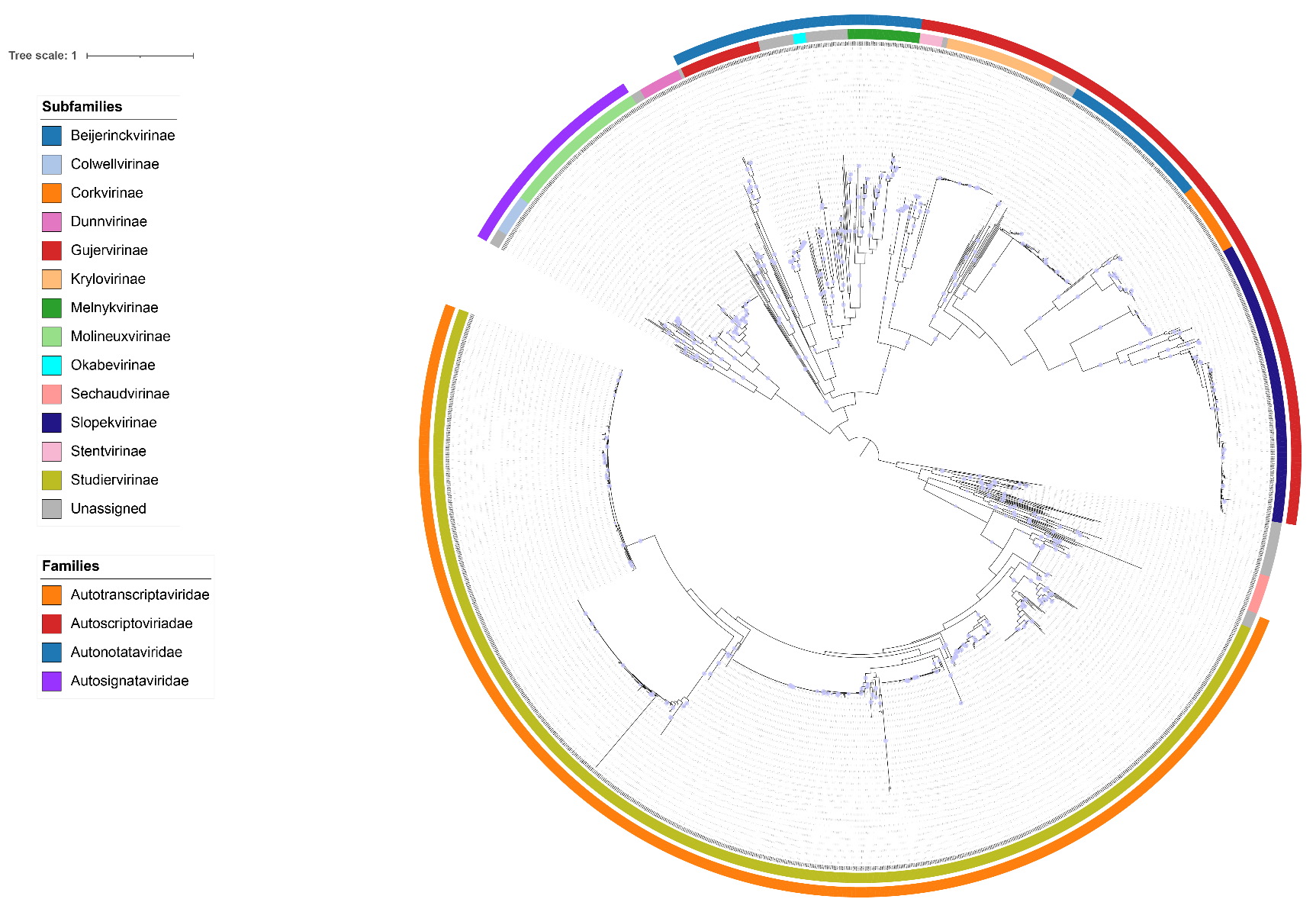
****

Figure 12. Single gene phylogeny of the tail tubular protein A. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

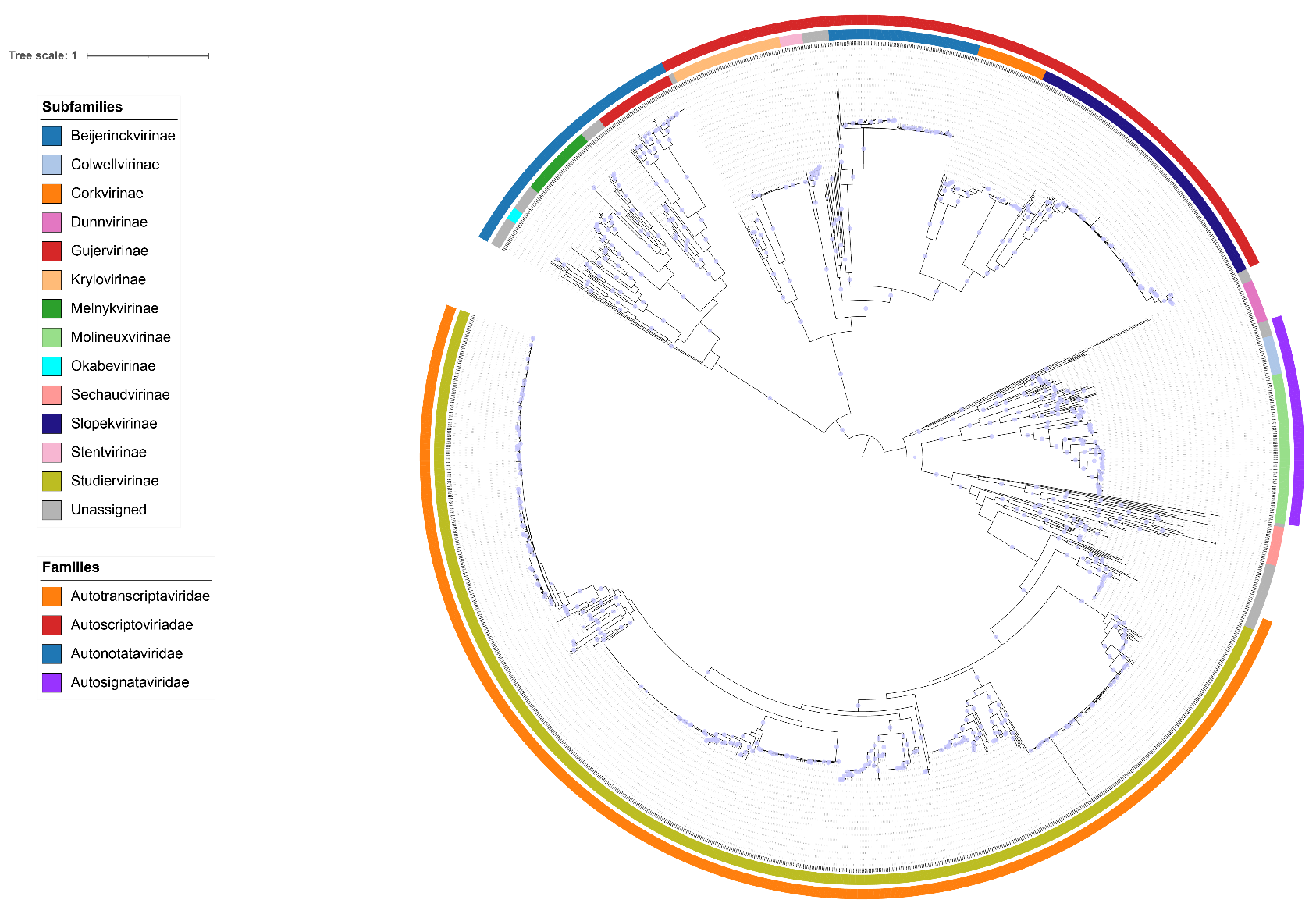
****

Figure 13. Single gene phylogeny of the tail tubular protein B. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers.

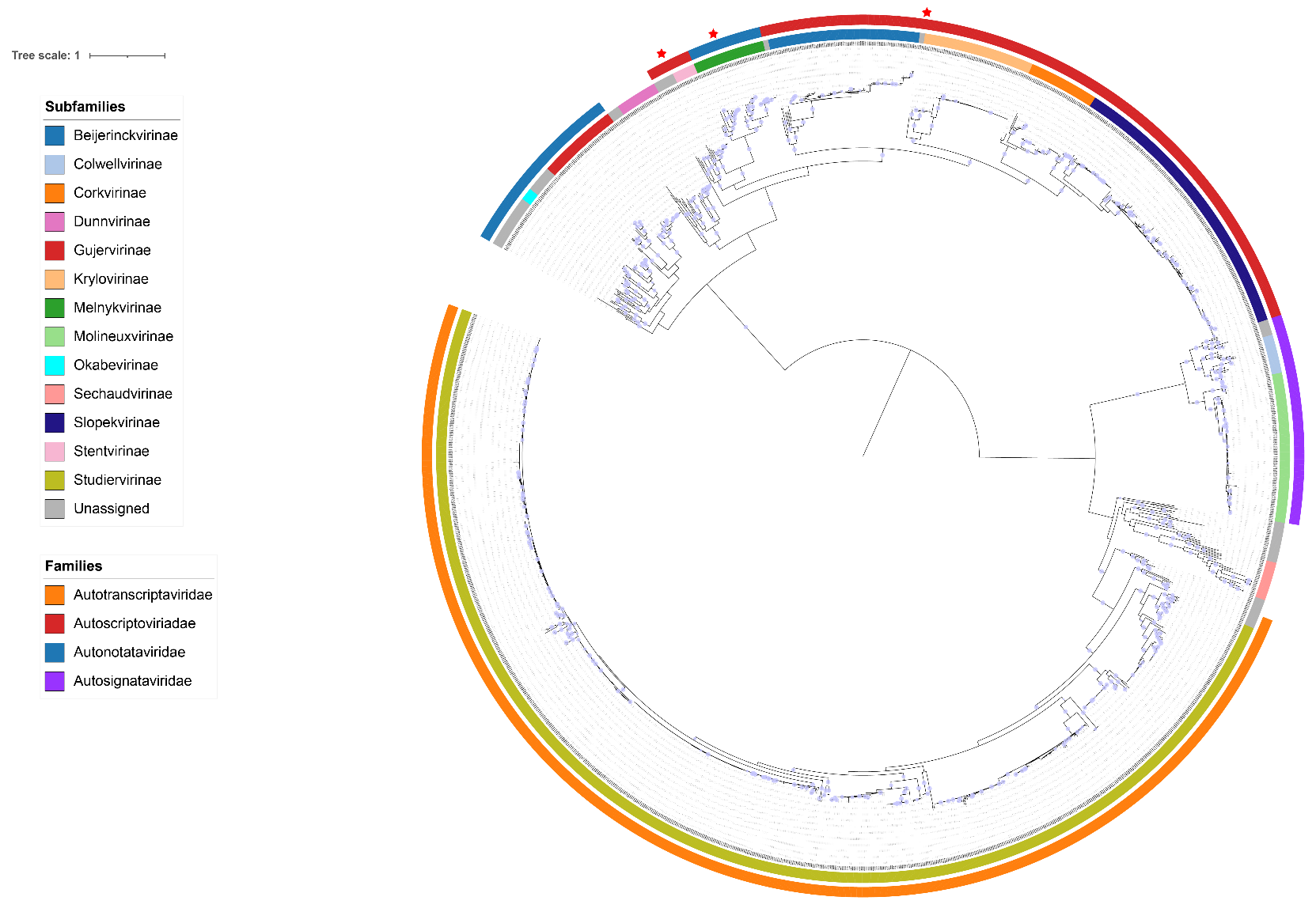
****

Figure 14. Single gene phylogeny of the DNA polymerase. Alignments were performed using MAFFT and trimmed using trimAl with a gap threshold of 0.5. The maximum likelihood tree was calculated using IQ-Tree2 with 1000 rapid bootstraps and SH-Alrt tests [14-16]. The tree is rooted at the midpoint and bootstraps ≥95% are shown as circles. The coloured rings indicate proposed subfamilies (inner ring) and families (outer ring). Node labels are INSDC accession numbers. Discrepancies between the single gene phylogeny and proposed classification are marked by a red star symbol.

Table 8. Derivation of names for ninety-two proposed new genera in the order *Autographivirales*.

|  |  |
| --- | --- |
| **Genus** | **Etymology** |
| *Actaeavirus* | Named after the Nereid Actaea in Greek mythology. |
| *Actinidiaevirus* | Name derived from the pathovar of *Pseudomonas syringae*, infected by phage Ep4. |
| *Aequorvirus* | Derived the Latin noun “aequor” for sea or ocean. |
| *Alfirinvirus* | Derived from the name of *Agrobacterium* phage Alfirin. |
| *Angmobvirus* | Name derived from “angered the mob”, as this virus was isolated from Baltimore inner harbor water. |
| *Aristophanesvirus* | Named after the first isolated phage of this genus, *Acinetobacter* virus Aristophanes. |
| *Armandvirus* | Named after the Center Armand-Frappier Sante Biotechnologie, Quebec, Canada where *Aeromonas* virus T7-Ah was isolated |
| *Atoyacvirus* | Named after the first isolated phage of this genus, *Aeromonas* virus Atoyac1. |
| *Axyvirus* | Named after the first isolated phage of this genus, *Achromobacter* virus vB\_AxyP\_19-32\_Axy09. |
| *Bamvirus* | Named after the first isolated phage of this genus, *Escherichia* virus vB\_Eco\_Bam. |
| *Benllochvirus* | Named after the street Carrer del Catedratic Agustin Escardino Benlloch where the Institute for Integrative Systems Biology (I2SysBio) is located. The first isolated phage of this genus, *Klebsiella* virus VLCpiA3a was isolated and characterized at I2SysBio. |
| *Bertilvirus* | Named after the first isolated phage of this genus, *Pseudomonas* virus Bertil. |
| *Boesrvirus* | Derived from the name of *Ralstonia* phage BOESR1. |
| *Bolekvirus* | Genus name proposed by Bujak et al., PMID: 35632757 |
| *Cankvirus* | Name derived from one of the symptoms caused by the host bacterium *Pseudomonas* syringae |
| *Catalonvirus* | Named after the Spanish region of Catalonia, the location of the University of Barcelona where bacteriophage Phi NF-1 was isolated and characterized. |
| *Ceskevirus* | Name derived from the city Ceske Bunejovice where the Institute of Plant Molecular Biology, Czechia is located and where *Xanthomonas* phage SB4 was characterized. ‘České’ translates as ‘Bohemian’. |
| *Chamilpavirus* | Named after the city where the Center for Genomic Sciences, Universidad s/n Col. Chamilpa is located. |
| *Conareevirus* | Name proposed by Bert Ely. |
| *Coryciavirus* | Name derived from the Greek nymph Corycia. |
| *Cotavirus* | Named after the first isolated phage of this genus, *Xylella* virus Cota. |
| *Cullenvirus* | Named after the Cullen building at the Baylor College of Medicine, where *Klebsiella* virus 6937 was isolated. |
| *Daeravirus* | Named after the Department of Agriculture, Environment and Rural Affairs (DAERA), Northern Ireland, which provided wastewater samples from which *Pectobacterium* virus MA13 was isolated. |
| *Daolivirus* | Named after the Daoli district in Heilongjiang province, China, where the Heilongjiang River Fisheries Research Institute is located. |
| *Dcimvirus* | Named after the phage DCM. |
| *Dishuivirus* | Name derives from the lake where *Synechococcus* phage DSL-LC07 was isolated. |
| *Dynamenevirus* | Named after the Nereid (sea nymph) Dynamene from Greek mythology. |
| *Ebriosvirus* | Named after the first isolated phage of this genus, *Escherichia* virus Ebrios. |
| *Euvesivirus* | Name derived from host of Xanthomonas phage SB3, *Xanthomonas* euvsicatoria. |
| *Fujianvirus* | Named after Fujian province in China, where the College of Aquaculture at JiMei University is located. |
| *Gansuvirus* | Named after the Chinese province of Gansu where the Lanzhou Veterinary Research Institute is located. |
| *Guangxivirus* | Name derived from the university where *Salmonella* phage PST\_H2 was isolated. |
| *Gundecimvirus* | Derived from the name of *Pseudomonas* phage G11. |
| *Hapakavirus* | Named after the river Hapaka that runs through Riga, Latvia. |
| *Hennigervirus* | Named after *Pseudomonas* phage Henniger. |
| *Hongshanvirus* | Named after the district of Wuhan where the State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University is located. |
| *Jelgvirus* | Named after the first isolated phage of this genus, *Aeromonas* virus JELG-KS1. |
| *Jeruvirus* | Derived from an abbreviation of Jerusalem, site of Hebrew University where *Providencia* virus PSTNGR1 was isolated. |
| *Jiaweivirus* | Named after *Mycobacterium* phage jiawei. |
| *Jimeivirus* | Derived from the road name where the Institute of Urban Environment, Chinese Academy of Sciences is located. |
| *Kuwvirus* | Derived from the name of *Paracoccus* phage ParKuw1. |
| *Linggongvirus* | Named after the road where the Dalian University of Technology, China, is located, where *Vibrio* virus vB\_VhaP\_VH-5 was isolated. |
| *Luzcentumvirus* | Derived from the name of *Pseudomonas* phage LUZ100. |
| *Maklayavirus* | Named after Miklukho-Maklaya street in Moscow, the location of the Laboratory of Molecular Bioengineering, IBCH where *Pectobacterium* virus Q19 was isolated. |
| *Mallvirus* | Name derives from isolation source, coastal sea water from Mallorca. |
| *Maltophvirus* | Name derived from the host bacterial species of phage BUCT609, *Stenotrophomonas maltophilia*. |
| *Mengvirus* | Derived from the name of *Nostoc* phage NMeng1. |
| *Micantvirus* | Named after the first isolated phage of this genus, Erwinia virus Micant. |
| *Mojovirus* | Name proposed by M. Raibey and R. W. Jackson, who isolated *Pseudomonas* virus MR5. |
| *Morelosvirus* | Named after the Mexican state of Morelos, where *Rhizobium* phage RHph\_I20 was characterized at the Centro de Ciencias Genomicas UNAM. |
| *Natansvirus* | Named after the species *Sphaerotilus natans*. |
| *Nerivirus* | Named using an acronym of the NUS Environmental Research Institute, Singapore, where the phage *Synechococcus* virus S-SRP01 was isolated. |
| *Nerthusvirus* | Named after the first isolated phage of this genus, *Pseudomonas* virus Nerthus. |
| *Njordvirus* | Named after the first isolated phage of this genus, *Pseudomonas* virus Njord. |
| *Oceanidvirus* | Derived from the name used to describe the innumerable daughters of the Titans Oceanus and Tethys in Greek mythology. |
| *Pakuvirus* | Named after the first isolated phage of this genus, *Burkholderia* virus Paku. |
| *Palovirus* | Named after the first isolated phage of this genus, *Rhizobium* phage Palo. |
| *Pasovirus* | Named after the first isolated phage of this genus, *Rhizobium* phage Paso. |
| *Pastovirus* | Named after the first isolated phage of this genus, *Rhizobium* virus Pasto. |
| *Paternavirus* | Named after the Spanish city of Paterna, where the Universitat de Valencia is located |
| *Pazvirus* | Named after the first isolated phage of this genus, *Xylella* virus Paz. |
| *Pfluvirus* | Named after the first isolated phage of this genus, *Pseudomonas* virus 22PfluR64PP. |
| *Ponderosavirus* | Named after first isolated phage of this genus, *Stenotrophomonas* virus Ponderosa. |
| *Proddevirus* | Named after first isolated phage of this genus, Desulfovibrio virus ProddE. |
| *Rambovirus* | Named after the first isolated phage of this genus, *Yersinia* virus vB\_YenP\_Rambo. |
| *Reminisvirus* | Named after first isolated phage of this genus, *Ralstonia* phage Reminis. |
| *Rindgevirus* | Name derived from location of Franklin Pierce University where *Escherichia* phage Tarrare was characterized. |
| *Rodentiumvirus* | Named after the host bacterial species of vB\_CroP\_CrRp3, *Citrobacter* rodentium. |
| *Roscoffvirus* | Named after the region of Roscoff in France, where the Laboratory of Integrative Biology of Marine Models is located. |
| *Sarmavirus* | Named after the first isolated phage of this genus, *Rahnella* virus Sarma103. |
| *Savitribaivirus* | Named after Savitribai Phule Pune University, India, where *Aeromonas* virus PS was isolated. |
| *Shangxiadianvirus* | Named after the road where the Fujian Agriculture and Forestry University is located. |
| *Smasvirus* | Name derived from an abbreviation of the host bacterium for viruses currently classified within the genus, *Stenotrophomonas maltophilia*. |
| *Snaubvirus* | Named after the first isolated phage of this genus, *Erwinia* virus pEp\_SNUABM\_03. |
| *Solymavirus* | Derived from the Greek word ‘Solyma’ which translates as Salem, referring to Jerusalem where Hebrew University is located. |
| *Songlingvirus* | Named after the Songling road, where the Qingdao Institute of Bioenergy and Biotechnology is located and where the first isolated phage of this genus, *Stappia* virus SI01, was characterized. |
| *Spiovirus* | Named after the Nereid Spio from Greek mythology. |
| *Sumtervirus* | Named after the street where the University of South Carolina is located. |
| *Tepoztlanvirus* | Named after Tepoztlan, a town located in the state of Morelos, Mexico, where *Rhizobium* phage RHph\_Y1\_20, the first phage of this genus was isolated. |
| *Thoosavirus* | Named after the Nereid Thoosa from Greek mythology. |
| *Tiranvirus* | Named after the island of Tiran in Red Sea as *Synechococcus* virus S-TIP28 was isolated from the Red Sea. |
| *Tolavirus* | Genus name proposed by Bujak et al., PMID: 35632757. |
| *Taurinorumvirus* | Named after the Roman name for the Italian city of Torino where the Insituto per la Protezione Sostenibile is located. |
| *Trelivelvirus* | Name derived from the Domesday book entry for Penryn. |
| *Triteiavirus* | Named after the daughter of the sea-god Triton, Triteia, in Greek mythology. |
| *Ulipvirus* | Named after the first isolated phage of this genus, *Klebsiella* virus K1-ULIP33. |
| *Unosvirus* | Named after the first isolated phage of this genus, *Pseudomonas* virus UNO-SLW1. |
| *Vistulavirus* | Named after the Vistula river than runs through Warsaw, the city where the Institute of Microbiology, University of Warsaw, Poland is located. |
| *Waldovirus* | Named after the first isolated phage of this genus, *Pseudomonas* phage Waldo5. |
| *Yautepecvirus* | Named after Yautepec, an area located in the state of Morelos, Mexico, where *Rhizobium* phage RHph\_X3\_9, the first phage of this genus was isolated. |
| *Yinyavirus* | Named after the river Yinya, the isolation source of *Aeromonas* virus Aer\_P220. |
| *Yuanmingyuanvirus* | Name derived from the road name where the Chinese Academy of Agricultural Sciences is located. |
| *Zoomievirus* | Named after the first isolated phage of this genus, *Erwinia* virus Zoomie. |