MICROBIAL GENOMICS

TECHNICAL RESOURCE
Pavan *et al.*, *Microbial Genomics* 2026;12:001632
DOI 10.1099/mgen.0.001632

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

CRESSENT: a bioinformatics toolkit to explore and improve ssDNA virus annotation

Ricardo R. Pavan[1,2,3], Matthew B. Sullivan[1,2,3,*] and Michael J. Tisza[4,*]

Abstract

ssDNA viruses are important components of diverse ecosystems; however, it remains challenging to systematically identify and classify them. This is partly due to their broad host range and resulting genomic diversity, structure and rapid evolutionary rates. In addition, distinguishing genuine ssDNA genomes from contaminating sequences in metagenomic datasets (e.g. from commercial kits) has been an unresolved issue for years. Here, we present **CRESSENT** (**CRESS**-DNA **E**xtended a**N**notation **T**oolkit), a comprehensive and modular bioinformatic pipeline focused on ssDNA virus 'genome-to-analysis' and annotation. The pipeline integrates multiple functionalities organized into several modules: sequence dereplication, decontamination, phylogenetic analysis, motif discovery, stem-loop structure prediction and recombination detection. Each module can be used independently or in combination with others, allowing researchers to customize their analysis workflow. With this tool, researchers can comprehensively and systematically include ssDNA viruses in their viromics workflows and facilitate comparative genomic studies, which are often limited to dsDNA viruses, therefore leaving behind a crucial component of the microbiome community under study. Benchmarking analyses demonstrated that CRESSENT efficiently processes ssDNA virus datasets of varying scales, completing small family-level analyses within minutes and moderate comparative genomics studies within hours using standard computing resources. Its modular, parallelized design ensures scalability and low memory usage, making it accessible to research groups with diverse computational capacities.

Impact Statement

The study of ssDNA viruses, especially circular rep-encoding ssDNA (CRESS-DNA) viruses, has expanded rapidly with the rise of metagenomic sequencing, yet researchers face persistent challenges in their annotation and comparative analysis due to their compact genomes, overlapping reading frames and limited functional references. This work addresses a critical gap by enabling more comprehensive and standardized characterization of ssDNA viruses. The outputs presented here represent an incremental but impactful step toward advancing ssDNA virus research, offering solutions that will benefit virologists, microbial ecologists and evolutionary biologists working with viral dark matter. This work enhances the interpretability and comparability of ssDNA virus genomes, ultimately supporting efforts to map the ecology and evolution of one of the most understudied components of the global virosphere.

Received 29 August 2025; Accepted 07 January 2026; Published 05 February 2026
Author affiliations: [1]Department of Microbiology, The Ohio State University, Columbus, Ohio, 43210, USA; [2]Centre of Microbiome Science, The Ohio State University, Columbus, Ohio, 43210, USA; [3]The Infectious Disease Institute, The Ohio State University, Columbus, Ohio, 43210, USA; [4]Department of Molecular Virology and Microbiology, The Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX, 77030, USA.
***Correspondence:** Matthew B. Sullivan, sullivan.948@osu.edu; Michael J. Tisza, michael.tisza@bcm.edu
Keywords: annotation; genomics; metagenomics; phylogenetic analysis; ssDNA; ssDNA virus.
Abbreviations: CRESSENT, CRESS-DNA Extended aNnotation Toolkit; CRUISE, CRiteria-based Uncovering of Iteron SEquences; MSL, Master Species List; vOTUs, viral operational taxonomic units.
All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary figures and supplementary material are available with the online version of this article.

DATA SUMMARY

CRESSENT is freely available online at https://github.com/ricrocha82/cressent. The documentation is available at https://cressent.readthedocs.io/en/latest/. Supplementary data are available at the Journal of General Virology online. The custom databases (contamination database and both family-level capsid and replication protein databases) are available for download from Zenodo under DOI: https://zenodo.org/records/15981951.

INTRODUCTION

ssDNA viruses are among the most abundant and diverse biological entities on Earth, yet they remain one of the least understood groups of viruses [1, 2]. The advent of metagenomic sequencing has led to a significant increase in the discovery of novel ssDNA viruses in diverse environments [2–4]. ssDNA viruses exhibit polyphyletic origins and have been repeatedly reshaped through gene exchange with plasmids, while the domain architectures of their Rep proteins vary markedly among bacterial, archaeal and eukaryotic hosts, collectively posing substantial challenges for comparative genomic analysis [5]. These challenges also include their small genome size (which can often lead to sequence removal in viromics workflows), rapid evolution rates, frequent recombination events and the presence of contaminating sequences in metagenomic datasets [6, 7]. Furthermore, the absence of universally conserved genes further complicates taxonomic classification and comparative genomics of ssDNA viruses [8, 9]. Due to these constraints, most studies examine a set of common genes within certain ssDNA groups, including Rep and Cap genes with conserved functions and structures [3, 10, 11]. Yet, these genes can be extremely divergent at the sequence level, and recombination frequently occurs between the Rep and Cap genes and, in some cases, within the Rep gene itself. Therefore, identification of these events involves phylogenetic analysis of the Cap gene, the Rep gene and/or the nuclease and helicase domains of the Rep gene [5, 12, 13], followed by tanglegrams to visualize recombination [13, 14]. Short motifs (e.g. Walkers A and B) within the nuclease and helicase domains can vary by clade, and these are often visualized as sequence logos [13]. Non-coding stem-loops and iterons are additional conserved structures essential for the replication of these viruses and can also be used to further delineate ssDNA virus lineages [14–16].

To address these challenges, we developed CRESSENT, a modular and comprehensive bioinformatic pipeline designed to analyse and annotate ssDNA virus sequences (Fig. 1). This tool consists of several modules, and each module can be used independently or in sequence, allowing researchers to customize their analysis workflow according to their specific research questions. Since the tool is modular, it allows for new modules to be added in the future. In this paper, we describe the functionality of each module and its integration into a cohesive analysis pipeline. With CRESSENT, researchers can enhance the efficiency and accuracy of ssDNA virus analysis and employ a standardized approach for comparative genomic studies of this important viral group.

The integration of multiple analysis methods provides researchers with a powerful toolkit for characterizing these important viral groups. The emphasis on data cleaning, visualization and flexible workflow design enhances the utility of CRESSENT for diverse

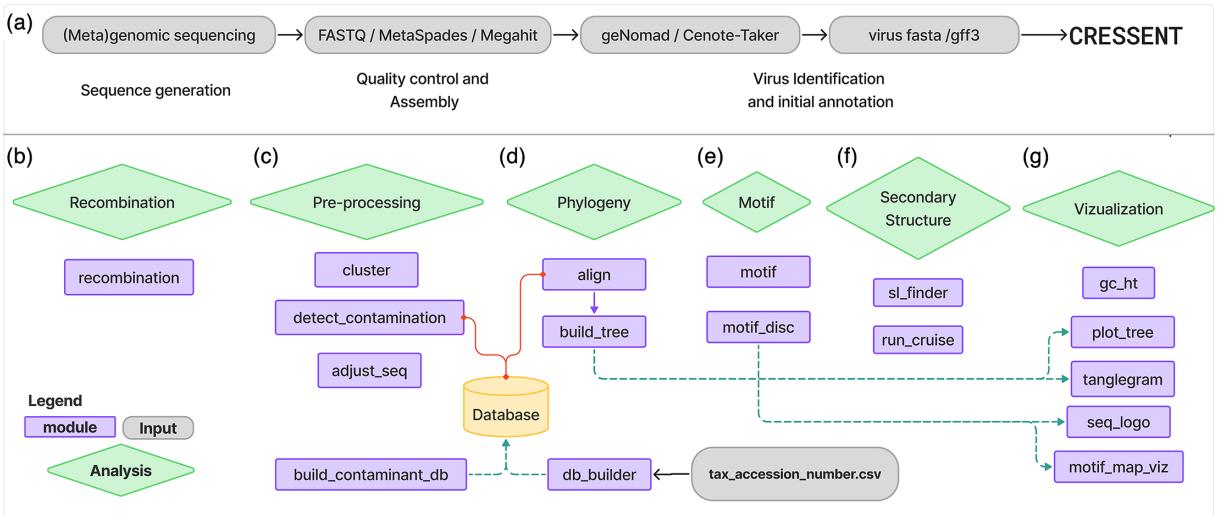

Fig. 1. Flow chart depicting the CRESSENT modules (blue boxes), analysis (green diamonds) and processes (purple boxes). Grey and green dashed arrows indicate potential pipelines. (a) The analysis starts with the commonly used tools for viral contig discovery and annotation. Then, the user can decide which module to use: (b) recombination analysis; (c) dereplication, decontamination or adjustment of the sequence based on determined nucleotide or protein sequence; (d) phylogenetic analysis; (e) motif analysis; (f) secondary structure detection; (g) visualization such as phylogenetic tree, tanglegram, GC content heatmap, sequence logo or motif map visualization.

research applications. We anticipate that this tool will contribute to a more standardized and reproducible approach to ssDNA virus analysis, facilitating comparative studies and advancing our understanding of viral diversity and evolution. Importantly, the modular architecture of this tool was specifically designed to facilitate continuous improvement and expansion. This extensibility ensures that the tool can evolve alongside the rapidly advancing field of viral metagenomics, providing a sustainable platform for ssDNA virus analysis for years to come.

CRESSENT MODULES

Existing virus discovery pipelines such as GeNomad [17], VirSorter2 [18] and Cenote-Taker3 [4] are highly effective at identifying viral contigs but offer limited post-processing capabilities specifically tailored to ssDNA viruses. CRESSENT addresses this gap by integrating widely used and validated bioinformatic tools into a single, modular and reproducible framework optimized for the evolutionary characteristics of ssDNA viruses. This design eliminates the need for users to manually locate, install and configure disparate tools, saving substantial time and ensuring that commonly required analyses in ssDNA viral genomics are performed in a consistent and standardized manner. The primary function of CRESSENT is therefore to serve as an auxiliary toolkit that refines the annotation and analysis of putative ssDNA viral sequences identified by upstream virus discovery programmes [4, 17, 18]. Moreover, its planned integration into the iVirus cyberinfrastructure [19] will facilitate seamless interoperability with other virus identification and annotation workflows.

The modular design of CRESSENT allows flexible integration of its components depending on specific research questions and dataset characteristics. Using the output of virus discovery tools as input (i.e. FASTA and GFF files), a typical workflow might begin with dereplication to reduce dataset complexity, followed by decontamination to remove potential laboratory contaminants. The resulting non-redundant and clean sequences can then undergo recombination detection, phylogenetic analysis and motif discovery, with stem-loop and iteron annotation further revealing replication-associated structures.

Although users may conceptually follow this logical sequence, CRESSENT does not enforce a fixed order of execution. Most modules are designed to accept standard FASTA or GFF3 files not only from other CRESSENT modules but also from external tools, allowing users to enter the workflow at any stage according to their data type or analytical goals. Even within a specific workflow such as phylogenetic reconstruction, iterative refinement is often necessary, as optimal alignment and trimming parameters can vary substantially between ssDNA viral families due to differences in genome organization and protein conservation.

The tool is written primarily in Python and R and is operated through a command-line interface. Each of the six modules is described in detail below.

RECOMBINATION DETECTION

Recombination is a major driver of genetic diversity and evolution in ssDNA viruses [20, 21]. The accurate detection of recombination events is therefore crucial for understanding viral evolution, taxonomy and epidemiology. The recombination module allows users to run specific or all methods for recombination detection and provides options for customization through a configuration file (Fig. 1b).

This module identifies recombination in nucleotide sequences using a suite of methods integrated within the OpenRDP (Recombination Detection Program) framework. The analysis incorporates multiple statistical and phylogenetic approaches to improve detection sensitivity and reliability [22]. The RDP method examines sequence triplets for recombination signals using a recursive segmentation algorithm, while 3Seq detects recombination based on phylogenetic incongruence among three sequences. GENECONV identifies gene conversion events by scanning for unusually long identical fragments shared between sequences. MaxChi and Chimaera both apply chi-square tests to detect breakpoints by comparing observed and expected substitutions across aligned sequences, with Chimaera using a more refined partitioning strategy. Bootscan assesses phylogenetic relationships along the sequence alignment by generating bootstrapped trees for sliding windows, highlighting regions of differing ancestry. SiScan extends this approach by calculating similarity scores across windows to detect potential recombination breakpoints. The methods implemented in this tool have been widely used in studies of viral recombination, including applications to circoviruses and parvoviruses, demonstrating their effectiveness across diverse ssDNA virus groups [23, 24].

The module accepts multiple-sequence nucleotide alignments (FASTA format) and produces a tab-delimited CSV table summarizing each detected event, including breakpoints, parental sequences, test statistics and the methods supporting it. Outputs also include timestamped log files documenting execution parameters and software versions to ensure reproducibility.

CRESSENT's recombination module uses the OpenRDP configuration, balancing detection sensitivity with runtime efficiency. Circular genome scanning is disabled by default (circular_genome=False) to prevent false breakpoints at contig termini, and Bonferroni correction ensures conservative significance control across multiple pairwise comparisons. Analytical *P*-values (num_permutations=0) accelerate analyses of large alignments, while events are reported when detected by at least one method (min_num_detecting_events=1). Each recombination algorithm operates with parameter sets tuned to compact, highly divergent

ssDNA genomes: RDP scans 30-nt windows, GENECONV permits short identical tracts and treats indels as polymorphisms and Bootscan employs 200 bp sliding windows (20 bp steps, 100 bootstraps) under a Jukes–Cantor model with ≥70% bootstrap support.

Complementary chi-square methods (MaxChi, Chimaera) use windows of 100–200 bp and require at least 60–70 variable sites per test, while SiScan performs similarity profiling with equivalent window sizes and permutation-based *P*-value estimation (1,100 permutations). These settings prioritize reproducibility and minimize false positives without oversmoothing short viral genomes. All parameters can be adjusted through an .ini file, enabling users to tighten detection criteria for high-confidence evolutionary analyses or to relax them for exploratory scanning across diverse ssDNA virus lineages.

DATA PRE-PROCESSING

The pre-processing step (Fig. 1c) involves three modules to cluster, clean or adjust the sequence for further analysis. First, the decontamination module addresses a critical issue in metagenomic studies: the presence of laboratory contaminants or 'kitome' sequences [25–27]. The user has full flexibility to define how stringent the decontamination process should be. Second, the user can use the adjust_seq module to rearrange the sequence, for example, to begin with conserved nonanucleotide sequences.

DEREPLICATION

The third method in data pre-processing, sequence clustering and dereplication, is widely used in viral metagenomics to manage dataset complexity and avoid bias from overrepresented sequences [7]. CRESSENT adheres to the current community standard for dereplication of viral contigs, as outlined in the MIUViG guidelines: a 95% ANI over at least 85% of the contig length (AF) [7]. This module utilizes a combination of tools, including anicalc.py and aniclust.py from CheckV [28] and BLAST [29], to calculate pairwise ANI and cluster sequences based on user-defined similarity thresholds. It outputs a vOTU catalogue with a designated cluster representative, the full set of clustered sequences, pairwise ANI values and supporting BLAST alignments. By removing redundancy while preserving true biological diversity, the module markedly reduces computational load for all downstream analyses.

DECONTAMINATION MODULE AND CONTAMINATION DATABASE

Laboratory contamination is a concern in microbiome metagenomic studies [25, 26, 30] and can lead to the misidentification of microbial taxa, inflate diversity estimates or result in erroneous biological interpretations if not properly controlled for through rigorous contamination-aware protocols. The systematic identification and removal of contaminant sequences is therefore essential for accurate analysis and interpretation of results.

It has been demonstrated that commercial reagents and extraction kits can contain viral DNA that may be misinterpreted as novel findings [31]. To mitigate this, a comprehensive database with 510 potential viral contaminants in laboratory reagents has been compiled [32]. Contaminants included four small circular virus-like genomes [25]. CRESS-like viruses were also found in laboratory reagents and may be constituents of the 'kitome'. For instance, the presence of CRESS-like viruses such as Parvovirus-like Hybrid Virus [33] and Rengasvirus [26] in negative control samples can be mistaken for novel or biologically relevant viruses, which highlights the need to cross-reference against known reagent-associated sequences to avoid erroneous conclusions in virome and pathogen discovery research.

To limit contamination, this module uses BLAST to compare input sequences against a curated database of known contaminants derived from five published studies [25, 26, 31–33]. The module produces decontaminated sequences, decontamination statistics and BLAST results for tracking potential contaminants. To balance the removal of contaminants while retaining genuine sequences, users can fine-tune the BLAST parameters, specifically the e-value (default: 1e-10), percent identity (default: 90) and alignment coverage (default: 50).

PHYLOGENETIC ANALYSIS, TREE VISUALIZATION AND CAP AND REP DATABASES

The phylogenetic analysis module (Fig. 1d) comprises three main components: sequence alignment, phylogenetic tree construction and tree visualization/annotation. Our implementation also includes user-friendly customization options, such as selecting family-level protein sequences from a ICTV-derived database focused on replication-associated (Rep) and capsid (Cap) proteins.

The alignment component uses MAFFT [34] for multiple sequence alignment and trimAl [35] for alignment trimming, while the tree construction component utilizes IQ-TREE2 for maximum likelihood phylogenetic inference [36]. MAFFT and trimAl have been used for viral sequence alignment due to their accuracy and efficiency, which is particularly important for divergent viral sequences [3, 8, 14]. IQ-TREE2 represents the state-of-the-art in maximum likelihood phylogenetic inference and has been used in viral phylogenomics [5]. Similar approaches have been used to investigate the evolutionary history of CRESS-DNA viruses, as well as to examine the diversity and evolution of ssDNA viruses in avian hosts [37, 38]. The visualization of phylogenetic

trees leverages specialized R packages, including ggtree [39], ape [40] and dendextend [41]. Additionally, when the tanglegram module is activated, the Robinson–Foulds distance is automatically calculated to quantify the topological dissimilarity between two phylogenetic trees. This metric provides a standardized way to assess how similar or different two trees are, which is essential for comparing phylogenetic reconstructions and evaluating the impact of different analytical methods or datasets [42].

In addition, a taxonomy-based database builder module is available for specific taxonomic levels (Fig. S1, Supplementary Material 1, available in the online Supplementary Material). The database is derived from the ICTV-recognized Master Species List (MSL) (MSL39.v4; https://ictv.global/sites/default/files/MSL/ICTV_Master_Species_List_2023_MSL39.v4.xlsx). The tool also supports user-generated databases and provides two dedicated modules for phylogenetic tree visualization: one for standard tree visualization and another for tanglegram plotting (Fig. 1g).

MOTIF ANALYSIS AND VISUALIZATION

Motif analysis is crucial for identifying functional elements in viral genomes, such as replication origins (ori), protein binding sites and conserved protein domains. In ssDNA viruses, conserved motifs within replication-associated proteins are especially informative, as they often reflect evolutionary constraints and mechanistic roles in rolling-circle replication. For example, motifs such as the HUH endonuclease domain and the superfamily 3 helicase domain are required for site-specific DNA cleavage and unwinding of the DNA strand, respectively. These functional domains are highly conserved across diverse ssDNA viruses, indicating strong purifying selection. As a result, their conservation across viral families can provide clues to the biochemical mechanisms of replication, robust markers for evolutionary comparisons and taxonomic classification [1, 3, 8, 43, 44].

The MOTIF modules (Fig. 1e) enable both pattern-based motif searching and *de novo* motif discovery in ssDNA viral sequences. The input can be nucleotide or protein sequences as FASTA files. Both modules produce several outputs, including motif positions, sequence logos of reference and user input motifs and genome maps to visualize the distribution of identified motifs.

For *de novo* motif discovery, the module integrates MEME for motif identification and optionally ScanProsite for scanning against known protein motifs. Both tools have been used for *de novo* motif discovery in viral genomics [45–47]. For example, tools such as MEME have been employed to explore conserved motifs in CRESS-DNA viruses, offering insights into their evolutionary relationships and classification [48, 49]. Similarly, motif identification platforms like ScanProsite have been applied to viral metagenomic sequences from avian hosts to detect functionally relevant sequence patterns [50]. The pattern-based searching functionality in our tool is particularly useful for identifying known functional motifs, such as the Walker A and Walker B motifs in Rep proteins, which are indicators of rolling-circle replication mechanisms common in ssDNA viruses [8, 44, 51]. For pattern-based searching, the module uses regex and the seqkit tool [52] to identify user-defined patterns and optionally split sequences at motif occurrences.

PUTATIVE STEM-LOOP AND ITERON ANNOTATION

DNA secondary structures typically form the origin of replication and serve as recognition sites for viral Rep proteins [14, 15, 53]. Stem-loop structures and iterons are essential for viral replication in many ssDNA viruses. As a well-studied example, the stem-loop structure in Faba bean necrotic yellows virus functions as the origin of replication, with a conserved nonanucleotide sequence within the loop being essential for initiating rolling circle replication [54]. Additionally, a set of iterative sequences (iterons) serves as specific binding sites for Rep proteins, and their spatial arrangement plays a crucial role in determining replication efficiency in *Geminiviridae*. These structures represent critical functional elements that are conserved across diverse ssDNA viral families despite high sequence variability in other genomic regions [55].

This module identifies and annotates critical noncoding secondary structures in ssDNA viral genomes: stem-loops and iterons (Fig. 1f). CRESSENT includes two modules for these analyses: StemLoop-Finder for identifying DNA hairpin structures and CRUISE (CRiteria-based Uncovering of Iteron SEquences) for detecting iteron repeats that serve as recognition sites for replication proteins [56, 57]. StemLoop-Finder uses the ViennaRNA library for predicting secondary structures and scoring potential stem-loops based on deviation from ideal stem and loop lengths. CRUISE identifies iteron sequences, which are typically found near the origin of replication in many ssDNA viruses.

VISUALIZATION AND OUTPUTS

Visualization components are integrated throughout the pipeline (Figs 1g and S2–S4 Supplementary Material 1), including tree visualization with ggtree, and sequence logo generation for motif representation. The visualization tools make use of established R packages such as ggtree [39], ggtreeextra [58] and ggseqlogo [59], providing publication-quality figures for downstream use. The output files from each module are designed to be compatible with the input requirements of subsequent modules, facilitating seamless integration. Also, they can be used as inputs for other tools not included in CRESSENT.

BENCHMARKING

To assess the computational performance and scalability of CRESSENT, we benchmarked its core modules using new metagenome-assembled genomes from two representative CRESS-DNA virus families: a small *Naryaviridae* dataset [60] containing three genomes (six proteins) and a *Genomoviridae* dataset [61] comprising 32 genomes (98 proteins). For each family, we independently evaluated analyses of the capsid (Cap) and replication-associated (Rep) proteins across four core computational modules: sequence clustering, multiple sequence alignment, phylogenetic tree inference and *de novo* motif discovery and annotation. All tests were conducted on a high-performance computing cluster (Red Hat Enterprise Linux 9.4) equipped with 40-core Intel Xeon processors and 187–375 GB RAM, using 40 threads for parallel execution.

The *Naryaviridae* dataset illustrated the pipeline's efficiency for small-scale analyses. Cap protein processing completed rapidly: clustering in 9 s (69 MB), alignment in <2 s (70 MB), motif discovery in 19 s (71 MB) and tree construction in 2.7 min (69 MB). Rep protein analyses showed similar performance: clustering (2 s), alignment (2 s), motif discovery (19 s), and tree construction (2.3 min) with all steps finishing within 3 min and <75 MB memory usage. For the *Genomoviridae* dataset, Cap protein clustering was completed in <3 s (87 MB), alignment required 11.4 min (988 MB), motif discovery 33 s (87 MB) and tree construction 5 h 24 min (1.37 GB). Rep protein analyses showed comparable clustering (6 s, 85 MB) and alignment (11.8 min, 1.44 GB) times, whereas tree inference was substantially more demanding (19 h 25 min, 1.96 GB), reflecting the higher sequence diversity and phylogenetic complexity of Rep proteins.

These results demonstrate that CRESSENT efficiently handles datasets ranging from small-scale studies to moderate-sized comparative genomics projects with reasonable computational requirements. Parallel implementation enables effective use of multicore systems, with high central processing unit (i.e. CPU) utilization observed during alignment and phylogenetic inference. Memory demands remain modest for most modules (typically 70–90 MB for clustering, alignment and motif discovery), with tree building constituting the primary computational bottleneck (70 MB–2 GB, depending on dataset size and protein diversity). Consequently, complete analyses can be executed on standard laboratory workstations for small datasets or scaled efficiently on high-performance computing platforms for larger projects, making CRESSENT accessible to laboratories with varied computational capacity.

Importantly, CRESSENT is designed for targeted analyses at the family, genus or species level, where phylogenetic signals remain informative for comparative genomics. Given the extensive sequence divergence among ssDNA viruses even within single genera [5, 13], joint analyses of multiple families or highly divergent lineages would markedly increase runtime and memory requirements, particularly during alignment and tree inference, while potentially reducing the biological interpretability of results due to alignment ambiguity and phylogenetic saturation (Fig. S1B–E).

Funding information
This work was supported by the U.S. National Science Foundation's Advances in Biological Informatics programme (2149505).

Acknowledgements
The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions
R.R.P.: conceptualization (equal), formal analysis (lead), investigation (lead), methodology (lead), software (lead), validation, visualization, writing – original draft and writing – review and editing (lead); M.B.S.: conceptualization (equal), funding acquisition (lead), project administration, resources (equal), supervision (supporting) and writing – review and editing (supporting); M.J.T.: conceptualization (equal), resources (equal), supervision (lead), investigation (supporting), methodology (supporting), software (supporting), validation, visualization and writing – review and editing (supporting).

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* 2012;157:1851–1871.
2. Tisza MJ, Pastrana DV, Welch NL, Stewart B, Peretti A, *et al*. Discovery of several thousand highly diverse circular DNA viruses. In: Kirkegaard K and Pride D (eds). *eLife*, vol. 9. 2020. https://doi.org/10.7554/eLife.51971
3. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, *et al*. *Cressdnaviricota*: a virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes. *J Virol* 2020;94:00582–20.
4. Tisza MJ, Belford AK, Domínguez-Huerta G, Bolduc B, Buck CB. Cenote-Taker 2 democratizes virus discovery and sequence annotation. *Virus Evol* 2021;7:veaa100.
5. Kazlauskas D, Varsani A, Koonin EV, Krupovic M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* 2019;10:3425.
6. Trubl G, Roux S, Solonenko N, Li Y-F, Bolduc B, *et al*. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* 2019;7:e7265.
7. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, *et al*. Minimum information about an uncultivated virus genome (miuvig). *Nat Biotechnol* 2019;37:29–37.
8. Kazlauskas D, Varsani A, Krupovic M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. *Viruses* 2018;10:187.
9. Krupovic M, Zhi N, Li J, Hu G, Koonin EV, *et al*. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. *Genome Biol Evol* 2015;7:993–1001.

10. Delwart E, Li L. Rapidly expanding genetic diversity and host range of the circoviridae viral family and other rep encoding small circular ssDNA genomes. *Virus Res* 2012;164:114–121.
11. Desingu PA, Nagarajan K. Genetic diversity and characterization of circular replication(Rep)-encoding single-stranded (CRESS) DNA viruses. *Microbiol Spectr* 2022.
12. Rosario K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Front microbiol* 2015;6.
13. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, *et al*. Evolutionary history of ssdna bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology* 2017;504:114.
14. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, *et al*. Unveiling crucivirus diversity by mining metagenomic data. *mBio* 2020;11:e01410-20.
15. Dai Z, Wang H, Xu J, Lu X, Ni P, *et al*. Unveiling the virome of wild birds: exploring CRESS-DNA viral dark matter. *Genome Biol Evol* 2024;16:evae206.
16. Torralba B, Blanc S, Michalakis Y. Reassortments in single-stranded DNA multipartite viruses: confronting expectations based on molecular constraints with field observations. *Virus Evol* 2024;10:veae010.
17. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, *et al*. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024;42:1303–1312.
18. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, *et al*. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;9:37.
19. Bolduc B, Zablocki O, Guo J, Zayed AA, Vik D, *et al*. iVirus 2.0: Cyberinfrastructure-supported tools and data to power DNA virus ecology. *ISME Commun* 2021;1:77.
20. Lefeuvre P, Lett JM, Varsani A, Martin DP. Widely conserved recombination patterns among single-stranded DNA viruses. *J Virol* 2009;83:2697–2707.
21. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, *et al*. Recombination in eukaryotic single stranded DNA viruses. *Viruses* 2011;3:1699–1738.
22. Martin DP, Williamson C, Posada D. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 2005;21:260–262.
23. Varsani A, Shepherd DN, Dent K, Monjane AL, Rybicki EP, *et al*. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virol J* 2009;6:36.
24. Fu X, Wang X, Ni B, Shen H, Wang H, *et al*. Recombination analysis based on the complete genome of bocavirus. *Virol J* 2011;8:182.
25. Duan J, Keeler E, McFarland A, Scott P, Collman RG, *et al*. The virome of the kitome: small circular virus-like genomes in laboratory reagents. *Microbiol Resour Announc* 2024;13:e0126123.
26. Keeler EL, Taylor LJ, Abbas A, Collman RG, Bushman FD. Rengasvirus, a circular replication associated protein-encoding single-stranded DNA virus-related genome that is a common contaminant in metagenomic data. *Microbiol Resour Announc* 2021;10.
27. Olomu IN, Pena-Cortes LC, Long RA, Vyas A, Krichevskiy O, *et al*. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC Microbiol* 2020;20:157.
28. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, *et al*. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;39:578–585.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
30. Olomu IN, Pena-Cortes LC, Long RA, Vyas A, Krichevskiy O, *et al*. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC Microbiol* 2020;20.
31. Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, *et al*. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect* 2019;25:1277–1285.
32. Porter AF, Cobbin J, Li CX, Eden JS, Holmes EC. Metagenomic identification of viral sequences in laboratory reagents. *Viruses* 2021;13:2122.
33. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, *et al*. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* 2013;87:11966–11977.
34. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
35. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
36. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, *et al*. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;37:1530–1534.
37. Chrzastek K, Kraberger S, Schmidlin K, Fontenele RS, Kulkarni A, *et al*. Diverse single-stranded DNA viruses identified in chicken buccal swabs. *Microorganisms* 2021;9:2602.
38. Olivo D, Khalifeh A, Custer JM, Kraberger S, Varsani A. Diverse small circular DNA viruses identified in an american wigeon fecal sample. *Microorganisms* 2024;12:196.
39. Xu S, Li L, Luo X, Chen M, Tang W, *et al*. Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data; 2022. https://onlinelibrary.wiley.com/doi/10.1002/imt2.56 [accessed 20 March 2025].
40. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;35:526–528.
41. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 2015;31:3718–3720.
42. Briand S, Dessimoz C, El-Mabrouk N, Lafond M, Lobinska G. A generalized robinson-foulds distance for labeled trees. *BMC Genomics* 2020;21:779.
43. Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, *et al*. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ* 2018;6:e5761.
44. Varsani A, Harrach B, Roumagnac P, Benkő M, Breitbart M, *et al*. 2024 taxonomy update for the family Circoviridae. *Arch Virol* 2024;169:176.
45. Bailey TL, Johnson J, Grant CE, Suite N. The MEME suite. *Nucleic Acids Res* 2015:w39–49.
46. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, *et al*. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 2006;34:W362–5.
47. Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics* 2002;1:107–108.
48. Requião RD, Carneiro RL, Moreira MH, Ribeiro-Alves M, Rossetto S, *et al*. Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains. *Sci Rep* 2020;10:5470.
49. Wang H-I, Chang C-H, Lin P-H, Fu H-C, Tang C, *et al*. Application of motif-based tools on evolutionary analysis of multipartite single-stranded DNA viruses. *PLoS One* 2013;8:e71565.
50. Vibin J, Chamings A, Klaassen M, Alexandersen S. Metagenomic characterisation of additional and novel avian viruses from Australian wild ducks. *Sci Rep* 2020;10.

51. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic circular rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. *Adv Virus Res* 2019;103:71–133.
52. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 2016;11:e0163962.
53. D'Souza DJ, Kool ET. Strong binding of single-stranded DNA by stem-loop oligonucleotides. *J Biomol Struct Dyn* 1992;10:141–152.
54. Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, *et al*. A single rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. *J Virol* 1999;73:10173–10182.
55. Bonnamy M, Blanc S, Michalakis Y. Replication mechanisms of circular ssDNA plant viruses and their potential implication in viral gene expression regulation. *mBio* 2023;14.
56. Jones A, Kasun GW, Stover J, Stedman KM, de la Higuera I. CRUISE, a tool for the detection of iterons in circular rep-encoding single-stranded DNA viruses. *Microbiol Resour Announc* 2023;12:e0112322.
57. Pratt AA, Torrance EL, Kasun GW, Stedman KM, de la Higuera I. StemLoop-Finder: a Tool for the detection of DNA hairpins with conserved motifs. *Microbiol Resour Announc* 2021;10:e0042421.
58. Xu S, Dai Z, Guo P, Fu X, Liu S, *et al*. ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol Biol Evol* 2021;38:4039–4042.
59. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 2017;33:3645–3647.
60. Zhang H, Fu Y, Cao C, Jiang H, Tang R, *et al*. Identification and characterization of novel CRESS-DNA viruses in the human respiratory tract. *Res Squ* 2025.
61. Leal Rodríguez C, Shah SA, Rasmussen MA, Thorsen J, Boulund U, *et al*. The infant gut virome is associated with preschool asthma risk independently of bacteria. *Nat Med* 2024;30:138–148.