**Abstract**

Viruses play a crucial role in agroecosystem functioning. However, few studies have examined the diversity of the soil virome, especially when it comes to RNA viruses. Despite the great progress in viral metagenomics and metatranscriptomics (metaviromics) toward RNA viruses characterization, soil RNA viruses' ecology is embryonic compared to DNA viruses. We currently lack a wet lab. method to accurately unhide the true soil viral diversity. To overcome this limitation, we developed dsRNA-based methods capitalizing on our expertise in soil RNA extraction and dsRNA extraction ported from studies of phyllosphere viral diversity. This proposed method detected both RNA and DNA viruses and is proven to capture a greater soil virus diversity than existing methods, virion-associated nucleic enrichment, and metaviromics. Indeed, using this method we detected 284 novel RNA-dependent RNA polymerases and expanded the diversity of Birnaviridae and Retroviridae viral families to agricultural soil, which, to our knowledge, have never been reported in such ecosystem. The dsRNA-based method is cost-effective in terms of affordability and requirements for data processing, facilitating large-scale and high-throughput soil sample processing to unlock the potential of the soil virome and its impact on biogeochemical processes (e.g. carbon and nutrient cycling). This method can also benefit future studies of viruses in complex environments, for example, to characterize RNA viruses in the human gut or aquatic environment where RNA viruses are less studied mainly because of technical limitations.

**Keywords:** RNA viruses, soil virome, RdRps evolution, dsRNA, Viromics, Metaviromics
Introduction

All cellular life forms are likely infected by viruses, which are the most ubiquitous and diverse biological entities on earth\(^1\). The soil is one of the highest viral-concentrated environments\(^2\). Through horizontal gene transfer, soil viruses influence microbial communities and act as a significant genetic resource that promotes biological diversity and evolution\(^3\).

The soil virome is important in ecology but remains inadequately characterized due to the difficulty in isolating viruses and their nucleic acids\(^2\). Soil micro-heterogeneity, matrix complexity, and organic inhibitors pose significant challenges to molecular biology techniques toward tapping soil viral diversity and abundance\(^4\). Metagenomics and microscopy-based studies have found DNA viruses to be the most abundant in soils\(^5,6,7,8\)\(^9\). However, current soil virus identification methods encompass a systematic bias toward prokaryotic viruses, particularly encapsulated viruses, to the detriment of eukaryotic viruses that lack capsid formation\(^10\). As a result, soil viral ecology research has mainly focused on DNA viruses of bacteria and archaea.

Given the possible host diversity in soil, it is expected that a large number of RNA viruses may exist. Most of the research on soil RNA viruses has focused on single hosts for important crops or crop diseases\(^15\)\(^16\). Generating and analyzing environmental RNA viral sequencing data is challenging due to technical limitations in extracting enough viral RNA from complex biotope samples and computationally finding viral genomic signature in large metatranscriptomic datasets\(^13\). To overcome these limitations associated capturing soil RNA viruses, a recent study has modified the VANA method to capture RNA viruses by eliminating the RNA digestion step and sequencing all extracted RNA including those from dominant organisms\(^10\). The authors were able to shown a divers community of soil RNA viruses, including viruses from bacteria, plants, fungi, vertebrates, and invertebrates, which give a novel insights on soil viral diversity.

Historically, two soil viral genetic material extraction methods has been widely used for detection of soil viruses through sequencing, Virion-Associated Nucleic Acids (VANA) and total RNA or DNA extractions. The VANA method is highly specific to viral species
detection, but encompass a bias toward DNA viruses\textsuperscript{11}. The second method is referred to metaviromics (metagenomics and metatranscriptomics applied to viruses)\textsuperscript{12}. The main challenges with use of soil metaviromics is that it relies on sequencing genetic materials dominated by organisms with larger genomes (e.g. bacteria, fungi etc.), making difficult to accurately identify viruses\textsuperscript{10, 13}. Furthermore, metaviromics is expensive, error-prone and time-consuming when trying to identify viruses\textsuperscript{14}. In addition, these methods for detecting soil viruses (VANA and metaviromics) can detect either DNA or RNA viruses at the same time and may introduce biases when characterizing the soil virus diversity.

Over the last five years, numerous metaviromics studies has help to expand soil RNA viral diversity and ecology\textsuperscript{11, 17, 18}. The majority of soil RNA viruses have been classified using the RNA-dependent RNA polymerase (RdRp) as a phylogenetic marker\textsuperscript{19}. This technique has been used to significantly increase the number of known RNA viral sequences in divers ecosystems\textsuperscript{20, 21, 22, 23, 24}. The RdRp plays a critical role in RNA virus genome replication and transcription, and is therefore widely used as a reference gene to estimate RNA virus phylogenies, forming the basis for the newly established classification scheme for RNA viruses\textsuperscript{25}. The RdRp domain sequences contain eight conserved amino acid motifs, with the motif A, B and C being highly characterized\textsuperscript{26}. The RdRp approach is a unique and novel way of studying RNA viruses, made possible by the development of advanced tools and databases that allow for efficient analysis of large datasets\textsuperscript{27}. In recent years, several new tools and databases have emerged and greatly facilitated the use of the RdRp approach in virology research, including the NeoRdRp database\textsuperscript{28} and the Muscle5 software package\textsuperscript{29}. These tools and databases enable the identification of conserved RdRp domains and motifs and the construction of robust phylogenetic trees\textsuperscript{30}. Despite these bioinformatics-based advanced tools, the effects of soil RNA viruses on soil ecological processes are still mostly unknown because of forementioned metaviromics limitations and the lack of viromic-based wet lab method to upscale virus ecology from soil ecosystems\textsuperscript{31}. Most, if not all, the viruses generate double-stranded RNA (dsRNA) as an intermediate during the viral replication process or as an erroneous product resulting from the bidirectional transcriptional readthrough\textsuperscript{32, 33, 34, 35}. Consequently all viral hosts have conserved mechanisms by which they sense viral infections through detection of dsRNA\textsuperscript{34}.  

\textcopyright CC-BY-NC-ND 4.0 International license available under a
Therefore, the dsRNA molecule can serve as a stable and resistant matrix for virus characterization, due to its thermodynamic stability and resistance to RNase destruction\(^\text{36}\), especially in situations where virus purification is challenging due to low titer, particle instability, or the presence of inhibitors. The extraction of dsRNA is a common approach widely used for the characterization of plant and fungi virome\(^\text{37, 38, 39, 40}\). Gaafar and Ziebell in 2020\(^\text{37}\) conducted a comparative study between viral particle extraction and dsRNA extraction in plants and concluded that the dsRNA-based approach should be prioritized due to its consistent ability to provide a more thorough understanding of the virome. However, to the best of our knowledge, there have been no published attempts to employ the dsRNA-based approach for soil virome characterization thus far.

In this study, a two-fold objective was pursued. Firstly, to enhance the detection and classification of viruses in organic and mineral soils through dsRNA-based methods, highlighting the significance of soil types and extraction method in analyzing RNA viral communities and their ecological roles. A cost-effective dsRNA-based method was developed and shown to expand the diversity of viral families in agricultural soil. Secondly, the investigation aimed to explore the soil virome's diversity through dsRNA-based approaches by identifying RdRp regions in assembled sequences. Different extraction methods showed distinct viral community compositions, with mineral soils having more eukaryotic viruses and organic soils having a more diverse community, including dsDNA viruses.

**Results**

For the sake of convenience and clarity, all the results (figures and tables) related to organic soil were put in the supplementary material section and can be easily consulted.

**Bioinformatics and data processing**

The sequencing results obtained from the three extraction methods (CCC-dsRNA, RPT-dsRNA, and VANA), were presented in table 1 and included the average number of reads, average number of bases, and average read length (after trimming) (Table 1 and Supplementary File 1). The results indicated that dsRNA-based methods produced both the
highest average number of reads and the highest average number of bases. On the other hand, VANA generated the longest average read length.

Table 1: Reads statistics for sequencing run of the three extraction methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. number of reads</th>
<th>Avg. number of bases</th>
<th>Avg. read length (after trimming)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-dsRNA¹</td>
<td>186,561</td>
<td>56,154,861</td>
<td>180.45</td>
</tr>
<tr>
<td>RPT-dsRNA²</td>
<td>218,943</td>
<td>65,901,730</td>
<td>205.81</td>
</tr>
<tr>
<td>VANA³</td>
<td>168,841</td>
<td>50,821,028</td>
<td>280.12</td>
</tr>
</tbody>
</table>

¹ Cellulose-column chromatography dsRNA extraction; ² RNeasy power soil total RNA kit dsRNA extraction; ³ virion associated nucleic acid approach.

To analyse our sequencing data, a new bioinformatics workflow was developed in Python language and compiled in a pipeline named SOVAP (Soil Virome Analysis Pipeline)⁴¹, which we described in the material and methods section. The pipeline comprises several modules that were designed to process the raw sequencing data and to identify, classify, and quantify the virome present in the samples (Fig. 1).
Figure 1: Overview of the SOVAP (soil virome analysis pipeline) workflow for viromics data processing and analysis. Module 1 includes quality checking and trimming of raw sequencing reads using Fastp. In module 2, contaminants are removed using Centrifuge. Clean reads are then assembled into contigs in module 3 using Megahit. In module 4, viral contigs are identified with the help of geNomad. In module 5 the abundance of viral contigs is estimated and redundant contigs are clustered. Module 6 involves the annotation and taxonomy assignment of viral contigs using Diamond and IMG/VR database.

In the second module of the SOVAP pipeline, which is dedicated to the decontamination step, we used Centrifuge software to filter out reads from non-viral sources. To achieve this, we created a custom database for Centrifuge which derived from SARs, Nematodes, Bacteria, Archaea, Fungi, Protozoa, and Human reference and representative genomes obtained from NCBI RefSeq (access date 16.03.2023) using Genome_updater v0.5.1 (https://github.com/pirovvc/genome_updater) (which includes 4,668 genomes)42. This custom database enables Centrifuge to accurately identify and remove non-viral reads from the virome datasets, resulting in more accurate and reliable viral detection and analysis. As part of the sixth module of the pipeline, a Diamond index of the Genbank viral database containing 58,201 genomes (access date 18.03.2023) is provided. This database is used in the last module when the IMG/VR database is not available or the pipeline is running in Diamond-Megan mode.

Viral taxonomy

The analysis of viral communities in mineral and organic soils using three different extraction methods yielded distinct taxonomic insights at various taxonomic ranks (Supplementary Files 2 and 3). The results showed that the VANA method displayed a viral communities dominated by DNA viruses, particularly Caudoviricetes, a type of tailed bacteriophages, which represented 92-94% of the TPM (Fig. 2 and Supplementary Fig. 1).

In mineral soil samples, the dsRNA-CCC and dsRNA-RPT extractions showed similar viral community compositions, with the majority of the viral contigs belonging to the Pisiviricota and Duplornaviricota phyla, which include both +ssRNA and dsRNA viruses. The Uroviricota phylum was also present in both extractions but was less abundant in the...
dsRNA methods than in the VANA extraction. The *Negarnaviricota* phylum, which comprises -ssRNA eukaryotic viruses, as well as *Kitrinoviricota* and *Lenaviricota*, which include +ssRNA eukaryotic and prokaryotic viruses, respectively, were also detected (Fig. 2, and Supplementary Fig. 3). Only the CCC-dsRNA extraction method was able to detect the presence of *Phixiviricota* phylum (non-enveloped ssDNA viruses) in this particular soil type (Fig. 2, and Supplementary Fig. 3). In contrast, the VANA extraction method yielded a different viral community composition in mineral soil samples, with the majority of the viral contigs belonging to the *Uroviricota* phylum, which includes tailed bacteriophages that infect bacteria. Additionally, *Nucleocytoviricota*, which includes dsDNA eukaryotic viruses, and *Cressdnaviricota*, which includes ssDNA viruses, were also present (Fig. 2, and Supplementary Fig. 3).

**Figure 2:** Comparison of taxonomic composition of viral communities in four mineral soil sample replicates extracted using three different methods: RNeasy Power Soil Total RNA kit.
(RPT-dsRNA), Cellulose-Column Chromatography (CCC-dsRNA), and the Virion Associated Nucleic Acid approach (VANA), at the taxonomic ranks of (a) realm, (b) kingdom, (c) phylum and (d) class. The stacked bar chart displays the percentage of assigned viral sequences for each taxonomic level.

At the family level, the dsRNA methods were dominated by Totiviridae and Partitiviridae, both including dsRNA viruses. Mitoviridae and Tombusviridae, which include +ssRNA eukaryotic RNA viruses and +ssRNA plant viruses, respectively, were present in both dsRNA extraction methods but were more abundant in the RPT-dsRNA extraction method. Both dsRNA methods also detected Retroviridae, a family of single-stranded RNA viruses that infect vertebrates. However, Birnaviridae, which includes dsRNA viruses, and Virgaviridae, which includes +ssRNA viruses, were only present in the RPT-dsRNA method (Fig. 3, and Supplementary Fig. 3).
Figure 3: Comparison of viral taxa abundance across four mineral soil sample replicates using three different extraction methods: RNeasy Power Soil Total RNA kit (RPT-dsRNA) and Cellulose-Column Chromatography (CCC-dsRNA), and the Virion Associated Nucleic Acid approach (VANA) at the (a) Order, (b) Family, (c) Genus and (d) Species levels. A heatmap shows the relative abundances determined by calculating the Transcripts per million (TPM) value of the reads counts of each viral taxa and their relative abundances in each sample replicate. The percentages of each viral taxa are indicated in the corresponding rectangles.

In organic soil samples, viruses of *Duplornaviricota* phylum were the most abundant in the RPT-dsRNA extraction method, while with the CCC-dsRNA method viruses of *Artverviricota* phylum were the most abundant, including all ssRNA-RT and dsDNA-RT viruses (Supplementary Fig. 1). The VANA approach revealed *Uroviricota* as the most abundant phylum and the presence of *Preplasmiviricota* phylum. At the family level when extracted using the VANA approach, the families *Tectiviridae*, *Straboviridae*, and *Demerecviridae*, which include dsDNA viruses, were more abundant. The *Inoviridae* family of ssDNA viruses and even *Mimiviridae* giant virus were observed not only in VANA but also in the CCC-dsRNA extraction (Supplementary Figs. 2 and 4). The RPT-dsRNA extraction method gave a more diverse viral community than the CCC-dsRNA method. Indeed, the RPT-dsRNA extraction method showed a different viral community composition, with the majority of the viral contigs belonging to the family *Cistroviridae*, dsRNA viruses that infect bacteria. Also unique families were observed using the RPT-dsRNA extraction method, including *Retroviridae* (+ssRNA viruses) and *Partiviridae* (dsRNA viruses) (Supplementary Figs. 2 and 4). The CCC-dsRNA extraction method revealed the presence of *Tombusviridae* (+ssRNA of plant viruses), *Phycodnaviridae* (dsDNA eukaryotic viruses), *Birnaviridae* and *Totiviridae* as well as phylum of *Phixiviricota* (non-enveloped ssDNA viruses) were detected with both RPT-dsRNA and CCC-dsRNA methods (Supplementary Figs. 2 and 4).

Viral diversity

*Alpha diversity.* The differences in the alpha diversity of viral DNA and RNA viruses using the three different virome enrichment methods: CCC-dsRNA, RPT-dsRNA, and VANA were also investigated. The richness, Shannon, and Simpson indices at different levels of taxonomic rank (realm, kingdom, phylum, class, order, family, genus, and species) showed...
significant (P < 0.05) differences in alpha diversity between the three extraction methods in both mineral and organic soil samples. Specifically, in mineral soil, the Simpson diversity index was significantly different at the realm (p = 0.015), kingdom (p = 0.015), phylum (p = 0.021), and class (p = 0.021) levels, while no significant differences (P > 0.05) were observed between the methods at the order, family, genus, and species levels (Fig. 4b and Supplementary Fig. 6b). The richness index was significantly different between the three extraction methods at the realm (p = 0.029), phylum (p = 0.045), class (p = 0.016), and order (p = 0.039) levels, with the CCC-dsRNA extraction method resulting in the highest index (Fig. 4a and Supplementary Fig. 6a). In organic soil, the Simpson index was significantly different between the three extraction methods, with the highest index recorded at the realm (p = 0.0097), kingdom (p = 0.0073), phylum (p = 0.0073), and class (p = 0.024) levels with the dsRNA extraction methods. At the family level (p = 0.025), the VANA extraction approach recorded the highest index (Supplementary Figs. 5b and 7b).

The Shannon index was significantly higher in mineral soil in both dsRNA-based methods at the realm (p = 0.015), phylum (p = 0.015), kingdom (p = 0.021), and class (p = 0.025) levels (Fig. 4c and Supplementary Fig. 6c), while it was higher in organic soil in the VANA method at the family (p = 0.023), genus (p = 0.009), and species (p = 0.0092) levels, and in the dsRNA extraction methods at the realm (p = 0.097), kingdom (p = 0.073), phylum (p = 0.018), and class (p = 0.018) levels (Supplementary Figs. 5c and 7c).
**Figure 4:** Alpha-diversity analysis of viral community composition in mineral soil using three extraction methods: RNeasy Power Soil Total RNA kit (RPT-dsRNA) and cellulose-column chromatography (CCC-dsRNA), and the Virion Associated Nucleic Acid approach (VANA). Violin plots show the Richness (A), Shannon (B), and Simpson (C) indices of all biological replicates of each method, calculated using Vegan v2.6-4 and Tidyverse packages. The analysis was performed at the realm, kingdom, phylum, class, order, family, genus, and species levels.
Beta diversity. Using the Bray-Curtis dissimilarity metric, Principal Coordinates Analysis (PCoA) was performed to compare viral communities in the two soil types extracted using the three different methods: CCC-dsRNA, RPT-dsRNA, and VANA. The beta diversity analysis showed distinct results for the mineral and organic soil samples. In mineral soil, the viral communities in the CCC-dsRNA and RPT-dsRNA methods clustered more closely compared to the viral community resulting from VANA extraction, indicating a more similar community composition. The viral community resulting from VANA method was clearly separated from the CCC-dsRNA and RPT-dsRNA samples at all taxonomical levels, suggesting a unique viral community (Fig. 5). In organic soil, the viral community resulting from the CCC-dsRNA and RPT-dsRNA methods not only clustered more closely, but interestingly there were shared viral community between the these two methods. In contrast, the viral community resulting from the VANA method was clearly separated from the viral community resulting from the dsRNA-based methods at the realm, kingdom, and phylum levels. The CCC-dsRNA, RPT-dsRNA, and VANA viral communities were also clustered separately at the class, order, and family levels, indicating distinct microbial communities at these taxonomic levels (Supplementary Fig. 8). The PERMANOVA results showed that the differences observed in the PCoA were significant (p < 0.05) at the different taxonomic levels for both soil types (Fig. 5 and Supplementary Fig. 8).
**Figure 5: Principal Coordinates Analysis (PCoA) plot of viral community composition in mineral soil, comparing the effectiveness of three extraction methods:** RNeasy Power Soil Total RNA kit (RPT-dsRNA), Cellulose-Column Chromatography (CCC-dsRNA), and Virion Associated Nucleic Acid approach (VANA), based on pairwise Bray-Curtis dissimilarities at each taxonomic rank. Each point on the plot represents a soil sample colored by the extraction method used. The p-value indicates the statistical significance of differences in viral community composition, as determined by a Permutational multivariate analysis of variance (PERMANOVA) test.

**Viral phylogenetic analysis.**

In both soil types, the dsRNA-based methods yielded an average of 5000-26000 viral contigs per sample through dsRNA extractions (Supplementary File 1). Therefore, we focused on the identification of novel contigs in *Riboviria* taxa candidates using RNA-dependent RNA polymerase (RdRp) as a phylogenetic marker and key RdRp motifs A, B, and C multiple sequence alignments analysis (Figs. 6 and 7). In total, we identified 3814 RdRps using functional RdRp protein motif annotations obtained from the RdRp-Scan workflow and database. Of these, 2944 were unique. Through multi-sequence alignment analysis of the conserved A, B, and C protein motifs of the RdRp domains (Fig. 6), we identified 284 potentially novel sequences that were classified into different phyla and
The Duplornaviricota (dsRNA viruses) phylum had the highest number of novel RdRps, with a total of 198 (Supplementary Fig. 9). We also identified 48 novel RdRps in the Pisuviricota phylum, which includes positive-sense single-stranded RNA and some double-stranded RNA viruses (Supplementary Fig. 10). In addition, we found 18 novel RdRps in the Kitrinoviricota phylum, which are positive-sense single-stranded RNA viruses that infect eukaryotes (Supplementary Fig. 11). We also identified 18 novel RdRps in the Lenarviricota phylum, which are positive-sense single-stranded RNA viruses that infect prokaryotes (Supplementary Fig. 12). Finally, we found two novel RdRps in the Birnaviridae family, which are dsRNA viruses (Supplementary Fig. 13).

Figure 6: Phylogenetic trees of RdRp sequences. The RdRp sequences extracted from mineral (inner ring, yellow) and organic (inner ring, pink) soil types using two dsRNA extraction methods, Cellulose-Column Chromatography (CCC-dsRNA) (outer ring, green) and RNeasy Power Soil Total RNA kit (RPT-dsRNA) (outer ring, black), were aligned with those reported in previous studies (WSH; Wolf, Starr, and Hillary, inner ring, white) and published by Hillary. The trees are divided into the RNA viral phyla, including Lenarviricota (a), Pisuviricota (b), Kitrinoviricota (c), Duplornaviricota (d), Negarnavirota (e), and Birnaviridae Family (f). The conserved motifs logos representing each Riboviria taxa that were extracted from the multiple sequence alignment results are shown in (g).
Figure 7: Phylogenetic trees of RdRp sequences. The RdRp sequences extracted from mineral (inner ring, yellow) and organic (inner ring, pink) soil types using two dsRNA extraction methods, Cellulose-Column Chromatography (CCC-dsRNA) (outer ring, green) and RNeasy Power Soil Total RNA kit (RPT-dsRNA) (outer ring, black), were aligned to the pre-built RdRp-Scan backbone trees published by Charon (inner ring white). The global RNA viral phylogeny is classified into the proposed RNA viral phyla, including (a) Lenarviricota, (b) Pisuviricota, (c) Kitrinoviricota, (d) Duplornaviricota, (e) Negarnavirota. (f) The diversity of RNA viruses across all identified taxa is visualized in a single phylogenetic tree using the RdRp-Scan whole tree backbone, providing a comprehensive view.

Discussion

Viruses are incredibly abundant on Earth, with estimates suggesting that there could be $10^{31}$ viruses$^{17, 18, 44, 45, 46}$. However, only $10^4$ viral species are currently listed in the International Committee on Taxonomy of Viruses (ICTV) official masters species list$^{47}$. In addition, when it comes to soil environments, our understanding of viral diversity and interactions is limited to selected groups of viral families largely dominated by DNA viruses$^{44}$. The difficulty and the cost in extracting soil RNA viruses has resulted in a slower
discovery rate compared to dsDNA viruses\textsuperscript{48, 13, 44}. In 2022, a novel metaviromics method (metagenomics and metatranscriptomics applied to viruses\textsuperscript{12}) was able to detect soil RNA viruses, but had missed important RNA virus taxa and still an expensive method for massive soil samples processing\textsuperscript{10}. The affordable and efficient viromics method proposed in the study is a known need for the scientific community to harness the soil viral diversity and fill the knowledge gaps on soil virus ecology, RNA viruses in particular.

The dsRNA-based methods captured a greater soil viral diversity than the standard method (VANA). When using the dsRNA-based methods compared to standard method, the alpha diversity was significantly higher at the four top ranks (real, kingdom, phylum and class) of virus taxonomy. The dsRNA-based methods in mineral soil yielded a significantly higher viral diversity and richness at the four and five top rank taxa, respectively. Indeed, with the VANA method the alpha diversity at all virus taxonomic levels was significantly lower (or similar) than dsRNA-based methods, except for organic soil where the alpha diversity obtained from the VANA method was significantly higher at the family and genus levels. However, the taxonomy analysis confirmed the known bias toward DNA viruses of the VANA method with 92-94\% of the TPM belonging to the class of Caudoviricetes, dsDNA viruses known as tailed bacteriophages. Regardless of the soil types, VANA method wasn’t able to detect RNA viruses of Riboviria realm and missed 50\% of all viral phyla compared to dsRNA-based methods. dsRNA-based methods were able to detect not only RNA viruses, but also DNA viruses except those from Prepsamiviricota phylum. In addition, the dsRNA-based method captured dsDNA-RT viruses (Artverviricota phylum) and non-enveloped ssDNA (Phixiviricota phylum), that VANA method missed. This result was supported by TruBl et al.\textsuperscript{5} who refined the VANA extraction-based approach to better capture both ssDNA and dsDNA viruses\textsuperscript{5}. Overall, these results were also reinforced through the beta diversity analysis, which indicated that dsRNA-based methods (CCC-dsRNA and RT-dsRNA) and VANA method captured distinct microbial communities.

In both soil types, the dsRNA-based methods yielded an average between 5000 - 26000 viral contigs per sample. In comparison, the only soil RNA-viromics study using deep sequencing to capture soil RNA viruses identified 3462 viral contigs\textsuperscript{10}. The primary dominant phyla in our dsRNA-based methods were Pisuviricota and Duplornaviricota.
This result differs from the soil RNA-viromics results, which identified a few number of dsRNA viruses of *Duplornaviricota* phylum\(^{10}\). Also, according to a recent study that compiled several datasets, *Lenarviricota* and *Pisuviricota* phylum were prevalent in soils\(^{14}\).

A possible explanation of this unexpected result may be that dsRNA method favor virus with dsRNA genome whereas metranscriptomic method favors virus with single-stranded RNA genome. Additionally, the denaturing step in the dsRNA method during cDNA synthesis may facilitate the reverse transcription of those sequences. This discrepancy need to be addressed though comparative study using mock sample sets. Contrary to VANA method, dsRNA-based methods detected viruses of *Partitiviridae* family in both soil types. *Partitiviridae* are bisegmented dsRNA viruses that can infect plants, fungi, and protozoa. These viruses only spread intracellularly during spore formation in fungi, which would explain why VANA method was not able to captured them\(^{49}\). In addition, the dsRNA method captured virus of *Retroviridae* taxon, a family of RNA viruses that possess a reverse transcriptase enzyme to produce DNA from RNA genome. To our knowledge, this virus family has never being reported in agriculture soil in the published literature even though viruses of this family are known to be widespread in nature\(^{50}\). The only reporting of this family in agricultural soil comes from a recent global study that used a novel bioinformatics approach within the new version of IMG/VR platform to search viruses in existing metaviromics datasets\(^{51}\).

Indeed, the use of metaviromics and high performance computing has increased our understanding of soil RNA virus diversity in five of the six main phyla of the *Riboviria* realm\(^{11, 18}\). Using dsRNA-based methods and RdRp-based phylogenetic analysis, an expanded fine-scale RNA viral diversity was revealed, with the six phyla containing RNA viruses within *Riboviria* viral realm detected compared to metaviromics studies\(^{10, 11, 19}\). We used two complementary approaches for detection of novel RdRps within the *Riboviria* phyla. We detected a total of 3814 RdRps whereas the two core soil metaviromics studies focusing on RdRps\(^{10, 11}\) detected 3422 and 3471 RdRps, respectively. Within our detected RdRp sequences, 2944 were unique according to the annotation of functional RdRps protein motifs derived from the recently published RdRp-Scan workflow and database\(^{52}\). Among these unique RdRps, 284 were putatively novels according to the multi sequences alignment analysis of the conserved A, B and C protein motifs of the RdRps domains.
These novels RdRps fell into Duplornaviricota phylum, Pisuviricota, Kitrinoviricota, Lenarviricota and undefined phyla. The two novel viral RdRps from the undefined phyla (incertae sedis) belong to Birnaviridae family, which, to our knowledge, have never been reported in agricultural soil before. Birnaviridae viruses are non-enveloped bi-segmented double-stranded (dsRNA) within Riboviria realm and do not employ a core particle in their replication cycle\textsuperscript{53}, thus would unlikely be captured through the recently published RNA-viromics, metaviromics or VANA methods\textsuperscript{10, 11, 18, 19}. Interestingly, we found these viruses in both soil types (organic and mineral), likely due to their host abundance, as the four genera of the Birnaviridae family infect insects, fishes, and birds\textsuperscript{54}. Furthermore, with dsRNA-based method we identified 37 viral contigs (with six unique RdRps) from the negative-sense RNA (-ssRNA) Negarnaviricota phylum, compared to only four viral contigs identified in the RNA-viromics study\textsuperscript{10}. These -ssRNA viruses are primarily lipid-enveloped, and may lack nucleocapsid proteins, making them difficult to detect using RNA-viromics approaches\textsuperscript{55}.

In conclusion, this study examined the effectiveness of two nucleic acid extraction methods, dsRNA and VANA, in characterizing the soil virome. The results indicated that the dsRNA method was more effective than the VANA approach in capturing the soil viral diversity. Moreover, the VANA and metaviromics methods are limited to detecting either DNA or RNA viruses, while the dsRNA method has proven to detect both DNA and RNA viruses. The study emphasizes the importance of selecting an appropriate extraction method based on the purpose of a given study. For researchers looking to identify DNA viruses, the VANA method may be suitable even though it missed non-enveloped ssDNA and dsDNA viruses, while those seeking RNA and DNA viruses should consider using the dsRNA method. Even though dsRNA-based method lacked of detecting DNA viruses of Preplasmiviricota phylum, it offers a better profile and coverage of the soil virome compared to existing methods (VANA and metaviromics). By breaking out the limitation associated with cost-effectiveness of using metaviromics to study soil RNA viruses, this work opens the door for intensive screening of soil RNA viruses diversity and ecology. It is a cost-effective method for massive screening of soil viruses and a proven tool to expand RNA viruses diversity. Furthermore, compare to existing soil-viromics methods,
sequencing dsRNA can give us an indication on virus that actively replicating in soil, because dsRNA is mainly produced as an intermediate during the viral replication process, except for viruses that has dsRNA as primary genome. The ds-RNA-based method can have broader impact beyond the soil environment and be adapted for in human gut or aquatic environments to characterize RNA viruses, which according to a recent studies are significantly less studied mostly because of their lack of stability and the difficulty of their identification in metaviromics sequencing\textsuperscript{13, 56, 57}. Moreover, this study highlights the known need of novel computational analysis methods for soil virus diversity studies and has developed the SOVAP workflow, which is a freely available open-access tool that can be adapted to study virus diversity and ecology.

Further steps would need a close comparison between the dsRNA-based method and RNA-based method (metaviromics approach) using the same samples to determine which method is preferable based on the targeted RNA virus groups and if combining both methods may provide a more comprehensive virome characterization. The developed dsRNA-based methods can be tested using nanopore sequencing to make more affordable.

### Methods

#### Sample collection

The samples used for this study were collected from two different experimental farms Agriculture and Agri-Food Canada's in Quebec, Canada. The first farm is a vineyard in mineral soil located in Frelighsburg and the four soil samples were collected at 5-15cm depths. The second farm is in organic soil in Sainte-Clotilde and the four samples were collected at 0-10 cm. The samples were properly stored at -20 °C until proceeding with nucleic acids extraction. Each of the eight samples was homogenized and divided in the three different groups corresponding to the three extraction methods, VANA, CCC-dsRNA and RPT-dsRNA, described in the following sections. The soil samples were submitted to AgroEnviroLab for analysis of the soil’s pH level and measurement of key exchangeable minerals, such as aluminum (Al), copper (Cu), zinc (Zn), manganese (Mn), and iron (Fe) (Supplementary Table 1).

**dsRNA extraction method using RNeasy Power Soil kit** (RPT-dsRNA extraction)
In this study, total RNA was extracted from the 8 soil samples, each containing 2 g, using the RNeasy Power Soil Total RNA kit (Qiagen, Hilden, Germany) as per the manufacturer's instructions. The resulting pellet was resuspended in 100 μl of RNase-free water. Enzymatic digestion was then carried out for all 8 samples using DNase I (Thermo scientific) at a final concentration of 10 U/μl and RNase T1 (Thermo scientific) at a final concentration of 10 U/μl for 30 minutes at 37°C. The purified dsRNA was obtained using the RNA mini elute kit (Qiagen, Hilden, Germany). The dsRNA was denatured by incubating it at 99 °C for 5 minutes, followed by immediate cooling on ice for 2 minutes. cDNA was then generated by incubating the purified dsRNA with random hexamers using the SuperScript IV cDNA Synthesis kit (Thermo scientific), followed by bead purification. Finally, the resulting cDNA was quantified using the Qubit dsDNA high-sensitivity kit (Fig. 8).

dsRNA extraction using cellulose-column chromatography (CCC-dsRNA extraction)

To isolate dsRNA using the cellulose-column chromatography method, total RNA was extracted from 2 and 1.5 grams of mineral and organic soil, respectively, using our optimized extraction protocol\(^{58}\). The extracted RNA was then subjected to a modified version of previously reported methods to isolate dsRNA\(^{59,60,61}\). Supernatant containing total RNA was mixed with absolute ethanol to obtain a solution of 16% ethanol. Furthermore, cellulose slurry was prepared using optimized chromatography buffer (50 mM Tris base, 100 mM NaCl, 16% ethanol, pH ~ 11) at a ratio of 200 mg cellulose per ml and mixed for 10 minutes to equilibrate the cellulose powder (Silmacell Cellulose, Type 101). 700 μl of the slurry was loaded into the PES centrifugal filter column (Sartorius, 0.8 μm) and centrifuged at 10,000 g for 10 seconds. Following that, the supernatant containing 16% ethanol was loaded on the column and centrifuged again. To further minimize humic contamination, four washes were performed using chromatography buffer to remove the majority of humic compounds from the column. To remove residual ethanol, the column was centrifuged at the same speed for 15 seconds prior to elution. The column was eluted with chromatography buffer (without ethanol, pH 8) followed by precipitation with 5 M LiCl final concentration and one volume of isopropanol for two hours. Upon dissolving the pellet in 20 μl ultrapure water and prior to cDNA synthesis, if traces of humic acid were detected, a final purification was conducted using the OneStep PCR Inhibitor Removal Kit.
After denaturing the dsRNA the cDNA was then generated as described in the previous section (RPT-dsRNA extraction). The resulting cDNA was then quantified using the Qubit dsDNA high-sensitivity kit.

**Virion associated Nucleic Acid extraction (VANA)**

To extract virion-associated nucleic acids, we used the protocol described in Santos-Medellin et al. Briefly, for each 50 g of soil sample, two 50 mL conical tubes were filled with 25 g of soil and 37.5 mL of 0.22 µm filtered AKC’ extraction buffer. The resulting mixture was homogenized, shaken on an orbital shaker at 400 RPM for 15 minutes, vortexed for 3 minutes, and centrifuged at 4,700 x g for 15 minutes. The supernatant was filtered through a 0.22 µm filter. The tube was then centrifuged at 32,000 x g for 3 hours at 4°C, and the resulting pellets containing the viral fraction were resuspended in 200 µl of ultrapure water. To remove free DNA, the resuspended pellets were treated with 30 units of DNase I (Thermo scientific) and 30 µl of 10X DNase buffer for 2 hours at room temperature, and the reaction was quenched with DNase stop solution. Nucleic acid was extracted using the Dneasy Power Soil Pro kit (Qiagen). The extracted DNA was quantified by employing a Qubit spectrofluorometry with Qubit dsDNA High Sensitivity kit. (Fig. 8).
Figure 8: Overview of the workflow of the three extraction methods from two soil types (mineral and organic). Three extraction methods were used to extract virus nucleic acids from the soil samples, which included two dsRNA extraction methods: the RNeasy Power Soil Total RNA kit (RPT-dsRNA) and Cellulose-Column Chromatography (CCC-dsRNA), and the Virion Associated Nucleic Acid approach (VANA) to extract viral DNA. After nucleic acid extraction and quantification, samples were prepared for library construction and quality control using Bioanalyzer, followed by sequencing on the Illumina MiSeq platform.

Library construction and sequencing

The libraries were prepared using the Nextera DNA library prep kit and 1 ng of input double-stranded cDNA. Subsequently, the Illumina Miseq platform sequencer was used to perform sequencing with the MiSeq Reagent Kits v3.

Virome data analysis. Adapters and low-quality reads were removed from raw sequencing data using Fastp v0.23.2. Subsequently, the resulting high-quality reads were decontaminated using Centrifuge v1.0.4 with a custom-built database. The clean reads were then assembled into contigs using MEGAHIT v1.2.9. The recently introduced geNomad v1.3.3 tool was utilized for the identification of viral contigs, which were then clustered using CD-HIT v4.8.1 for further analysis. The abundance of the viral contigs was estimated using BWA v0.7.17-r1188, and an in-house script was used to calculate abundance values for each contig. The clustered viral contigs were annotated using DIAMOND (blastx) v2.1.4 with the in-house indexed IMG/VR v4 database. The resulting output was visualized using R v4.2.0, and the taxonomy was plotted with Tidyverse v1.3.2 sub-packages for each taxonomy rank.

The alpha-diversity analysis, including species richness, Shannon, and Simpson indices, was calculated and visualized using the Vegan v2.6-4 and Tidyverse packages, respectively. Detailed steps and codes for the entire process are available in the GitHub repository (https://github.com/poursalavati/SOVAP_Soil_dsRNA). Additionally, a similar analysis was conducted by concatenating all biological replicates of each method and analyzing them together, following the same methodology as described earlier. Moreover, beta-diversity analysis was conducted to evaluate the dissimilarity and similarity of viral communities at each taxonomy rank between the viromics methods using...
the Bray-Curtis distance matrix, and principal coordinate analysis (PCoA) was employed to visualize the results\textsuperscript{74}. Permutational multivariate analysis of variance (PERMANOVA) was used to determine the statistical significance of the observed differences.

**Phylogenetic analysis.** Diamond blastp and RdRp-Scan datasets\textsuperscript{27} were used to search for RdRp genes in the identified viral proteins from geNomad (> 70 amino acids). A second analysis of the resulting hits was conducted using HMMSearch v3.3.2\textsuperscript{76} with the RdRp-Scan Hidden Markov model (HMM) profile database to improve the accuracy of putative RdRp sequences. Then contigs with E values < 0.001 and scores >50 were clustered with CD-Hit to remove redundant sequences (with 95% average amino acid identity). The putative candidates were then categorized into different taxa and aligned using MAFFT v7.515\textsuperscript{77} and Clustal Omega v1.2.4\textsuperscript{78} to the RdRp-Scan pre-built backbone trees, as well as previously reported RdRp sequences from the Wolf\textsuperscript{79}, Starr\textsuperscript{20}, and Hillary\textsuperscript{80} studies (WSH) for each individual taxa, respectively. Trees were generated with FastTree (MP, Multi-threaded) v2.1.11\textsuperscript{81} and visualized with iTOL v6.7.3\textsuperscript{82}. We used a multi-step approach to re-analyze RdRp candidates and identify novel sequences. First, we identified three key RdRp motifs (A, B, and C) using a motif database\textsuperscript{27} and CLC Genomics Workbench v22 (QIAGEN). Next, we performed multiple sequence alignment (MSA) with MegAlign Pro v17.4.1 (DNAStar Lasergene software) to identify conserved regions and potential variations among the candidates of each Riboviria taxa. Finally, we visualized the MSA results and conserved motifs logos using Inkscape v1.2.1 (https://inkscape.org). This approach allowed us to effectively analyze the candidates and identify important functional motifs in potential novel instances.

**Data availability**

The sequencing data supporting this research have been deposited in the NCBI Sequence Read Archive (SRA) repository under BioProject accession number PRJNA948674. Detailed bioinformatics workflow (SOVAP scripts), along with the generated databases, are available at https://github.com/poursalavati/SOVAP, https://zenodo.org/record/7758200 and https://zenodo.org/record/7747335, respectively. Custom R scripts for downstream data analysis, along with results of abundance, alignments, phylogenetic
trees, and identified RdRp sequences from all datasets are accessible on the following GitHub repository: https://github.com/poursalavati/SOVAP_Soil_dsRNA.

Acknowledgments

The authors gratefully acknowledge the administrative support from the CRD of Saint-Jean-sur-Richelieu team, including Vicky Toussaint and Mélanie Cadieux teams. We would also like to express our gratitude to Joel Lafond-Lapalme for his valuable support in maintaining the bioinformatics system. Furthermore, we acknowledge Eric Courcesne and his team for their support and management of the experimental vineyard.

Funding

This research was funded by Agriculture and Agri-Food Canada under the Ensuring water and soil resources sustainability (J-002302) and ecosystem productivity and resiliency (J-001792) programs.

References


42. Tange O. GNU Parallel 20221122 (‘Херсон’).). Zenodo (2022).

43. Hillary LS, Adriaenssens EM, Jones DL, McDonald JE. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels.) (2022).


