Benchmarking of virome metagenomic analysis approaches using a large, 60+ members, viral synthetic community

Deborah Schönegger, Oumaima Moubset, Paolo Margaria, Wulf Menzel, Stephan Winter, Philippe Roumagnac, Armelle Marais, Thierry Candresse

Benchmarking of virome metagenomic analysis approaches using a large, 60+ members, viral synthetic community

Deborah Schönegger[1], Oumaima Moubset[2,3], Paolo Margaria[4], Wulf Menzel[4], Stephan Winter[4], Philippe Roumagnac[2,3], Armelle Marais[1] and Thierry Candresse[1*]

[1]*Univ. Bordeaux, INRAE, UMR BFP, CS20032, 33882, Villenave d'Ornon Cedex, France.*

[2] *CIRAD, UMR PHIM, 34090 Montpellier, France*

[3] *PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France*

[4] *Plant Virus Department, Leibniz-Institute DSMZ, 38124 Braunschweig, Germany*

*** Author for correspondence:** thierry.candresse@inrae.fr Tel : (+33) 557 12 23 89

Abstract word count: 248 words

Text word count: 7561 words (Introduction-Materials and Methods-Results-Discussion)

Tables and Figures: 5 Tables, 4 Figures, 7 supplementary Tables and Figures

Running title: Synthetic viral community for virome benchmarking

Abstract

In contrast to microbial metagenomics, there has still been only limited efforts to benchmark virome analysis approaches performance in terms of faithfulness to community structure and of completeness of virome description. While natural communities are more readily accessible, synthetic communities assembled using well characterized isolates allow more accurate performance evaluation. Starting from authenticated, quality-controlled reference isolates from the DSMZ Plant Virus Collection, we have assembled synthetic communities of varying complexity up to a highly complex community of 72 viral agents (115 viral molecules) comprising isolates from 21 viral families and 61 genera. These communities were then analyzed using two approaches frequently used in ecology-oriented plant virus metagenomics: a virion-associated nucleic acids (VANA) based strategy and a highly purified double-stranded RNAs (dsRNA) based one. The results obtained allowed to compare diagnostic sensitivity of these two approaches for groups of viruses and satellites with different genome types and confirmed that the dsRNA-based approach provides a more complete representation of the RNA virome. For viromes of low to medium complexity, VANA however appears a reasonable alternative and would be the preferred choice, in particular if analysis of DNA viruses is of importance. They also allowed to identify several important parameters and to propose hypotheses to explain differences in performance, in particular differences in the imbalance in the representation of individual viruses using each approach. Remarkably, these analyses highlight a strong direct relationship between the completeness of virome description and sample sequencing depth which should prove useful in further virome analysis efforts.

Importance

We report here efforts to benchmark performance of two widespread approaches for virome analysis, which target either virion-associated nucleic acids (VANA) or highly purified double-stranded RNAs (dsRNA). This was achieved using synthetic communities of varying complexity levels, up to a highly complex community of 72 viral agents (115 viral molecules) comprising isolates from 21 families and 61 genera of plant viruses. The results obtained confirm that the dsRNA-based approach provides a more complete representation of the RNA virome, in particular for high complexity ones. For viromes of low to medium complexity, VANA however appears a reasonable alternative and would be the preferred choice if analysis of DNA viruses is of importance. Several parameters impacting performance were identified as well as a direct relationship between the completeness of virome description and sample sequencing depth. The strategy, results and tools used here should prove useful in a range of virome analysis efforts.

Keywords: virome, VANA, dsRNA, synthetic community, metagenome, double-stranded RNA, high-throughput sequencing

INTRODUCTION

Significant advances in the development of molecular methods have been made in the last decades, including innovative sequencing technologies based on DNA/RNA approaches such as targeted (RT-)PCR or non-targeted High-Throughput Sequencing (HTS). HTS, also known as next generation sequencing (NGS), enables high-speed, high-throughput sequencing of native DNA/RNA or amplified DNA, generating enormous amounts of sequencing data. These developments led to major advances in the field of metagenomics, i.e. the sequencing of the entire genetic material of a sample, and to a new understanding of microbial diversity [1, 2]. Viral metagenomics has revealed the immense diversity and ubiquity of viruses in nature and thus revolutionized our vision of these biological agents [1, 3-8]. Specifically, these metagenomics studies have revealed that virus sequence data available in public databases are biased toward human viruses or viruses of anthropological significance, with e.g. influenza-like viruses found in fish and amphibian hosts [9] or more than 75% of the plant virus species characterized up to 2006 having been isolated from crops [10]. These findings, together with reports on viruses associated with hosts different from those known for the vast majority of their relatives, such as flavi-like viruses found in plants [11, 12], have raised novel questions about virus-hosts co-divergence or host switching.

In plant virology in particular, advances in the development of viral metagenome analyses have been of great importance in terms of early detection of known viruses and discovery of novel plant viruses [4, 7, 13-14], as more than half of emerging diseases in plants are thought to be caused by viruses [15]. HTS has a huge potential in plant virus diagnostics because it allows to picture the complete phytosanitary status of a plant and to differentiate between virus variants that may contribute differentially to disease etiology [14]. For example, in a metagenomic analysis of sour cherry showing symptoms of Shirofugen stunt disease (SSD), a divergent isolate of little cherry virus 1 (LChV1) was identified in the absence of any other viral agent,

63 suggesting that LChV1 could be responsible for the SSD disease [16]. However, metagenomics
64 approaches have also revealed that plants are often infected by more than one virus [17],
65 complicating the unravelling of the etiology of plant viral diseases.

66 HTS has also renewed the link between classical plant virology and ecology [4, 18]. Viromes
67 identified from both cultivated and uncultivated plant populations enabled the study of
68 ecological processes such as the movement of viruses between different host reservoirs, the
69 effects of management practices or of the anthropological simplification of ecosystems [19-23].

70 For the efficient characterization of complex plant-associated viromes, there is generally a need
71 to enrich viral sequences and conversely reduce the amount of host plant sequences that are
72 generated. Different target nucleic acid populations have been used for virome studies but,
73 coupled with the virus enrichment constraint, the most widely used approaches have targeted
74 virion-associated nucleic acids (VANA) or double stranded RNAs (dsRNAs) [4, 7]. For single
75 plant samples or low complexity samples, the use of total RNA or small interfering RNA
76 (siRNA) sequencing are considered the most universal and straightforward options [24, 25] but
77 when the viromes of entire plant communities are analyzed from complex plant pools, VANA
78 or dsRNAs enrichment methods are generally preferred [4, 7, 19, 21, 26]. A huge number of
79 bioinformatic tools are available for HTS data analysis and have been, together with nucleic
80 acid preparation strategies, extensively reviewed [13, 27-28]. The choice of a specific viral
81 enrichment method or bioinformatic pipeline depends on the experimental objectives. Even
82 though there have been some efforts towards performance comparisons of different virome
83 analysis approaches [29, 30], there is a need to better benchmark them and assess their
84 respective efficiency at providing a faithful and comprehensive description of complex viromes,
85 without introducing biases. In a virus discovery study on single quarantine plants, VANA was
86 shown to assemble longer contigs compared to siRNA for a novel DNA mastrevirus [31], while
87 in a study investigating the virome of native plants in Oklahoma, more viral Operational

Taxonomy Units (OTUs) could be detected with dsRNA compared to VANA [26]. Ma *et al.* [32] provided a more comprehensive comparison of these two approaches using the natural viral communities present in complex plant pools from managed and unmanaged sites. The authors found significant differences with more viral contigs and, on average, longer contigs assembled from libraries prepared from dsRNA. With regard to viral richness, more OTUs were detected by the dsRNA approach compared to the VANA one. However, most DNA viruses were only detected using VANA.

Standardization is fundamental for the reliable representation of microbiome/virome in metagenomic studies and is challenged by the rapid development of sequencing platforms, protocols and bioinformatic pipelines [33]. Benchmarking is a powerful tool to provide standards that can be used to compare and evaluate the performance of the different steps required in metagenomic studies, including target nucleic acids population extraction, library preparation, sequencing (and sequencing platform) and finally bioinformatics sequence analysis. In this context, benchmarking studies in metagenomics are often based on mock communities that are microbial assemblages of known composition which can be used to compare the actual *vs* the expected performance of a process. Besides the use of actual empirical phytoviromes [32], the use of synthetic communities could therefore provide a more precise and detailed benchmarking of HTS-based virome description strategies. Bacterial and fungal mock communities have thus been developed and used to compare the performance of different sequencing platforms, e.g. short read Illumina or long read PacBio SMRT sequencing [34-36]. In recent years, viral mock communities have also been developed, especially in the medical and clinical field, to benchmark protocols in human virome studies. For example, the nucleic acid preparation step for the virome analysis of fecal samples was optimized using a combination of both viral and bacterial mock communities [37]. In another study, the bias introduced by viral enrichment or random amplification were assessed using a DNA virus mock

community [38]. Viral synthetic communities have also been used to benchmark library preparation approaches in environmental [39] and insect [40] virome studies. However, the use of synthetic communities in plant virome studies is lagging behind. So far, the only study using a defined mix of plant viruses to assess different nucleic acid preparation protocols was performed by Gafaar and Ziebell [30]. This study revealed a better performance of enriched dsRNAs as compared to ribodepleted total RNA or siRNAs for virus detection. However, only low complexity synthetic communities have been used so far, whereas most of the viral metagenomes associated with natural plant communities are composed of a complex and diverse mixture of DNA and RNA viruses that are studied from pooled plant samples. In the present work, we used a total of 22 synthetic plant virus communities of varying degrees of complexity to compare the diagnostic performance of VANA and dsRNA-based approaches for virome description and analyzed how this performance is affected by sequencing depth and other parameters. In parallel, a first attempt at contrasting the performance of VANA and dsRNA approaches with those of RNASeq was conducted, using synthetic datasets assembled *in silico* from single-isolate RNASeq data.

MATERIALS AND METHODS

Mock viral communities design

A list of 61 different viruses (assigned to 59 different genera from 18 different families plus one unassigned virus) was selected among those kept in collection and available at the Leibniz-Institute DSMZ - German Collection of Microorganisms and Cell Cultures (Braunschweig, Germany), taking into consideration three main criteria: (i) maximizing viral diversity by including viruses with all genome types (ssDNA, dsDNA-RT, dsRNA, +ssRNA, -ssRNA), (ii) including (with one exception) only a single representative virus per viral genus and (iii) selecting viruses/isolates for which a complete or near complete genomic sequence is available. In some cases, these genomic sequences had been determined previously, while in other cases they were developed specifically in the frame of efforts to further improve the characterization

of isolates distributed by the DSMZ through the EU-funded EVA-Global initiative (https://www.european-virus-archive.com/). Quality controlled samples were obtained from the DSMZ in the form of infected, lyophilized plant material in vacuum-sealed vials. The complete list of the isolates used, together with their properties and the propagation host in which they were provided, are given in Table 1.

Initial low complexity pools were generated by assembling 30 mg of virus-infected samples into 12 viral communities comprising five viruses each (150 mg of plant material each) and containing at least one virus with a genome type different from +ssRNA (Supplementary Table S1). Pea enation mosaic virus was counted as one virus, when it is in fact a co-infection of pea enation mosaic virus 1 (*Enamovirus*) and pea enation mosaic virus 2 (*Umbravirus*). Stepwise combinations of these five viruses mock communities were then assembled to create communities of increasing degrees of complexity (Supplementary Fig. S1), yielding a total of 22 communities with complexity ranging from five to 60 viruses.

Double-stranded RNA extraction

Double-stranded RNAs were purified from pooled samples according to [41] with some minor modifications. Briefly, instead of 75 mg, 150 mg dried plant material (representing a pool of five plants, Supplementary Table S1) was used as starting material and buffer volumes increased proportionally. Plants were ground in liquid nitrogen until a fine powder was obtained which was then mixed with the phenol-extraction buffer. Following gentle agitation for 30 min and centrifugation, the supernatant was decanted and half of it directly further processed, while the other half was used for the stepwise gradual assembly of pairs of communities used to generate more complex viral communities. In this way, six communities of 10 viruses each, then three communities of 20 viruses and finally a single community of 60 viruses could be assembled. Between each step, assembled samples were vortexed for at least 30s for optimal homogenization. A detailed scheme of the pooling strategy to form communities of different

complexities is shown in Supplementary Figure S1. Irrespective of its complexity, a supernatant volume corresponding to an initial input of 75 mg of plant sample was thus obtained and further processed as per the protocol of Marais *et al.* [41] which involves two rounds of CC41 cellulose (Whatman) chromatography followed by a nuclease treatment (DNase RQ1 plus RNaseA under high salt conditions) to remove any remaining host DNA and single-stranded RNA. A negative extraction control using only buffer was systematically included. Purified dsRNAs were finally converted to cDNA and randomly amplified while simultaneously adding MID tags [41-42].

VANA extraction

VANA extractions were performed on pools of five viruses similarly prepared as for dsRNA, using the protocol of François *et al.* [42] with minor modifications. Briefly, 150 mg of lyophilized plant material (representing a pool of five plants, Supplementary Table S1) were ground in Hank's buffered salt solution (HBSS) (1:10) with four metal beads within a grinding machine (Fastprep 24, MP Biomedicals). Following two centrifugation steps (4000g for 5 min at 4°C and 8000g at 4°C for 3 min), the supernatants were split and used in the same stepwise assembly of more complex communities as for the dsRNA approach (Supplementary Figure S1). A negative, buffer only, extraction control was systematically included. Each of the thus generated samples, representing different degrees of community complexity, was filtered through a 0.45µm filter and centrifuged at 148,000g for 2.5 hours at 4°C to concentrate the virus particles. Unprotected nucleic acids were eliminated by DNase and RNase treatment at 37°C for 1.5 hours. Viral RNA and DNA were then isolated using the NucleoSpin Virus kit (Macherey Nagel, Hoerdt, France), using only 80 µl of sample in the first lysis step and omitting the addition of proteinase K. Extracted RNAs were transformed to cDNA using Superscript III reverse transcriptase (ThermoFisher Scientific/Invitrogen), cDNAs were further purified with the QIAquick PCR purification Kit (Qiagen, Courtaboeuf, France) and a complementary strand was synthesized using the Klenow fragment of DNA polymerase I. Finally, a random PCR

amplification adding barcoded dodeca-linkers and corresponding MID primers during reverse transcription and PCR, respectively was performed [42].

Illumina sequencing

PCR products from all communities analyzed using the dsRNA and VANA procedures were finally purified using the MinElute PCR purification kit (Qiagen) and equimolar quantities of amplification products were sent to Illumina sequencing in multiplexed format (2 × 150 bp) on two lanes (one for VANA and one for dsRNA, respectively) on a NovaSeq 6000 system at the GetPlaGe platform (GenoToul INRAE Toulouse, France).

Generation of synthetic datasets for viral communities using single-isolate RNASeq data

For all but one of the viral isolates used to build the synthetic communities, available single-isolate ribodepleted RNASeq datasets (Leibniz-Institute DSMZ) were used to reconstruct *in silico* datasets corresponding to the different communities with reads number and average reads length paralleling those from the VANA and dsRNA datasets. These reconstructed datasets, mimicking the analysis of the various communities by RNASeq, were analyzed in parallel to those generated by the VAN A and dsRNA approaches.

HTS data analysis

Sequencing reads were imported into CLC Genomics Workbench v. 21.0.3. (CLC-GW, Qiagen) and adapters were removed from reads followed by trimming on quality and length using default settings and a minimum read length of 60 nucleotides (nt). Final trimmed reads were on average 111-113 nt long for the various datasets. Datasets were normalized by resampling at varying depth as needed, using the random reads sampling tool in CLC-GW.

To analyze virus detection performance as a function of contig size, *de novo* assembly was performed with CLC-GW (word size, 50; bubble size, 300) using various minimum contig lengths (125, 175, 250, 350, 500, 1000 nt). In order to identify viruses possibly present in the

samples used, in addition to the expected reference viruses, contigs were annotated by a BlastX analysis [43] against the viral RefSeq portion of the non-redundant (nr/nt) NCBI GenBank database. For the additional viruses thus identified, a genomic scaffold was reconstructed and extended by repeated rounds of residual reads mapping using CLC-GW, thus yielding near complete genome sequences that were used as reference for the relevant virus (Table 2). In a few cases, these assemblies were considered too incomplete and the closest complete genomic sequence in GenBank was selected as reference sequence (Table 2).

In order to determine virus detection performance, unassembled reads or *de novo* assembled contigs were mapped against the reference genome segment(s) for each virus (Tables 1 and 2) using very stringent mapping parameters (length fraction 100%, minimal similarity fraction 90%) in CLC-GW. In order to take into account inter-sample crosstalk due to index jumping [44-45], a threshold of positive detection was computed for each viral molecule by calculating the average plus 3 standard deviations (SD) of background virus reads observed in libraries generated from communities that did not contain the corresponding virus. Assuming a normal distribution of background reads, the use of such a positivity threshold would provide a <1% risk of reporting a false positive detection (https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule).

Comparison of parameters (number, average length) for *de novo* assembled viral contigs obtained from VANA and dsRNA datasets normalized at different sequencing depths were performed with five resampling repeats at each depth. Statistically significant differences were identified using a two-sample t-test.

Data availability

Trimmed sequencing reads for all viral communities analyzed by dsRNA or VANA approaches are available from the French *Recherche Data Gouv* multidisciplinary repository at https://doi.org/10.57745/42WNRJ. The normalized 10M reads dsRNA or VANA datasets

generated using the 60-viruses community have also been made available together with the community composition and the complete or near complete reference genomic sequences used are also available from the same repository at https://doi.org/10.57745/T4UYPC.

RESULTS

Viruses or virus-like agents identified from viral communities HTS data

The analysis of reads from both the VANA and dsRNA approaches for all communities revealed the presence of all expected viruses, although a few viruses were only represented by a limited number of reads or were only detected using one of the two approaches. Overall, only lettuce ring necrosis virus turned out to be fully absent from VANA reads, while banana bunchy top virus was only represented by a single dsRNA read. It should also be noted that not all viruses could be detected in all the communities of different complexity in which they were expected.

In addition to the expected 61 viruses, evidence for the presence in some communities of additional viruses or virus-like agents was obtained through the BlastX indexing of *de novo* assembled contigs from the low complexity, 5-viruses communities. A total of 11 unexpected agents were thus identified (Table 2). These include three linear ssRNA satellites associated with the helper virus isolates included in the communities [turnip crinkle satellite F (TCVsatRNA F), pea enation mosaic satellite RNA (PEMVsatRNA) and strawberry latent ringspot virus satellite RNA (SLRSVsatRNA)], latent viruses associated with the propagation hosts used [Hordeum vulgare endornavirus (HvEV), maize-associated totivirus (MaTV), maize-associated totivirus2 (MaTV-2) and Chenopodium quinoa mitovirus 1 (CqMV1)], as well as viruses in coinfection with some of the viral isolates used [poinsettia mosaic virus (PnMV), tobacco mosaic virus (TMV), turnip yellows virus (TuYV) and maize streak Réunion virus (MSRV)] (Table 2). Taken together, these agents represent three additional viral families,

263 for a total of 21 viral families (plus satellites) used for the assembly of communities. For these
264 additional agents, either a nearly complete genome was reconstructed from sequencing reads
265 and used as the mapping reference or the closest full genome sequence in GenBank was used
266 for further mapping analyses (Table 2). For all other viral isolates included in the communities,
267 complete or nearly complete genomic sequences were available (Table 1).
268 While the communities of varying complexities analyzed here will be referred to as 5-viruses,
269 10-viruses, 20-viruses and 60-viruses, it should be kept in mind that the real number of viruses
270 present in a given community might be slightly different because of (i) the presence of one or
271 more of the additional viruses and (ii) the counting of pea enation mosaic virus as one virus
272 when it is in fact a co-infection of pea enation mosaic virus 1 (*Enamovirus*) and pea enation
273 mosaic virus 2 (*Umbravirus*).

274 **Read mapping analysis of VANA and dsRNA datasets for the communities of various**
275 **complexities**
276 To be able to compare results between low and high complexity communities, all datasets were
277 normalised by randomly subsampling 120K cleaned reads, the depth of the 5-viruses
278 community with the lowest number of reads. To address the issue of inter-sample crosstalk
279 caused by index jumping [44-45] a threshold of positive detection was computed for each viral
280 molecule by calculating the average + 3 standard deviations (SD) of background reads in
281 libraries generated from communities that did not contain the corresponding virus. Assuming a
282 normal distribution of crosstalk reads numbers, this strategy ensures that the probability of
283 having a mapped reads number higher than the threshold by chance (false positive detection) is
284 lower than 1%.

285 In general, the proportion of viral reads in both VANA and dsRNA datasets was high (64-89%)
286 and was slightly affected by community complexity, with a general trend to reach higher values
287 when analysing more complex communities (Figure 1A). The proportion of viral reads in the

dsRNA datasets were slightly higher than in the corresponding VANA datasets, with the strongest differential observed for lower complexity communities of five and 10 viruses (64-65% viral reads as compared to 79-82%, Figure 1A). In contrast, the average proportion of viral reads in RNASeq datasets for individual virus isolates following ribodepletion was 19.6% but with a very large standard deviation of 26.1%.

Using the 12 communities of five viruses and a sequencing depth of 120K reads, 67 viruses were detected with both VANA and dsRNA approaches (with detection of reads for at least one genomic molecule considered as positive detection for a virus with a multipartite genome), out of the total of 72 viruses or virus-like agents present in the 12 communities analyzed (93.1%). However, VANA yielded reads for all six DNA viruses used (100%), while dsRNA yielded reads for only three of them (50%). Conversely, VANA yielded reads for 61 of the 66 RNA viruses or satellites (92.4%), when dsRNA yielded reads for 64 of them (97.0%) (Figure 1B). As expected, and previously reported, the performance of VANA is thus superior for DNA viruses but that of dsRNA slightly superior for RNA viruses. Using the datasets reconstructed from single plant RNASeq data, an overall rate of detection of 97.2% of the 71 viruses was obtained (no RNASeq data was available for one of the isolates used, which was therefore excluded from all computations).

The impact of increasing community complexity is reflected by the diminishing number of viruses detected at an equal sequencing effort of 120K reads. The performance of VANA gradually deteriorated, with detection decreasing from 61 RNA viruses detected to 58 (10-viruses communities) and then to 52 (20-viruses communities) to reach only 34 RNA viruses detected (51.5%) in the most complex community (Figure 1B). The same pattern was observed for DNA viruses, with all six DNA viruses detected in the 10- and 20-viruses communities but only one detected when analysing the 60-viruses community. In the case of the dsRNA approach, performance was marginally reduced for the 10- and 20-viruses communities (65 and

63 RNA viruses detected, respectively) and less affected than for the VANA approach for the most complex community, with still 57 of 66 RNA viruses detected (86.4%) (Figure 1B). Remarkably, performance was the least affected for the RNASeq approach using reconstructed communities data, with still 65 viruses (91.5%) detected for the most complex community (5/6 DNA viruses and 59/65 RNA viruses, or 90.7%).

If trying to compensate for community complexity by proportionally increasing the sequencing effort for more complex communities, the erosion in performance is less important for VANA, with still 57 of 66 RNA viruses detected for the 60-virus community (86.4%) and five of the six DNA viruses (83.3%) at a 1.44 M reads depth (12 x 120K). The performance of dsRNA, on the other hand, is no longer impaired, as all 66 RNA viruses (100%) were detected for the most complex community (result not shown). Similarly, the performance of RNASeq was no longer substantially impacted, with all DNA viruses and all but one RNA viruses detected.

The stronger degradation of VANA performance as community complexity increases, correlates with a more uneven distribution of read numbers between viruses and the stronger dominance of a few viruses, in particular turnip yellow mosaic virus (TYMV). In the 60-viruses community VANA dataset, TYMV represented 67% of the reads while the corresponding value for the dsRNA dataset was only 28%. As shown in Figure 2, even if spanning a 5 to 6 logs scale, the percentage of reads for each virus in the total datasets tends to be more evenly distributed between viruses in the dsRNA dataset than in the VANA dataset for the 60-viruses community. By contrast and excluding a single sample showing extremely low viral reads numbers, the variation in the proportion of viral reads in individual viral isolates analyzed by RNASeq showed much less variability as it remained within a 3 logs range of variation.

Although allowing to compare the performance of the VANA and dsRNA approaches, these analyses based on the mapping of reads against cognate reference genomes do not mimic the situation in metagenomic studies, in which a high proportion of viruses are expected to be novel

and for which therefore no suitable reference genome is available. We therefore analyzed the performance of these two approaches following the *de novo* assembly of reads into contigs, which is known to reduce the proportion of un-annotated "dark matter" [46].

Impact of minimal contig length on the number of detected viruses

We first evaluated the impact of the minimal contig length on the number of detected viruses using the most complex community of 60-viruses and deep datasets normalized at 10 M reads. As expected, and shown in Figure 3, the number of detected viruses decreased as minimal contig length increased. The pattern observed for RNA viruses is similarly observed for DNA viruses. The dsRNA approach consistently detected more RNA viruses than the VANA one, irrespective of the minimal contig length used, but the difference increased as minimal contig length increased. Using the shortest, 125 nt contig length, VANA identified 54 of the 66 RNA viruses or satellites present in the community (81.8%), while dsRNA identified 63 of them (95.5%) (Figure 3). The corresponding values for DNA viruses are respectively 4/6 (66.7%) and 3/6 (50%).

On the other hand, the coverage of the detected viruses (fraction of the target molecules represented in contigs) was much less affected by minimal contig length. While being relatively stable for the dsRNA approach, for which it varied between 66.5% and 74.9% with no clear trend, it showed a tendency to increase with contig length for the VANA approach, from 50.2% (>125 nt contigs) to 76.7% (>1,000 nt contigs) (Supplementary Figure S2).

For further analyses, an intermediate 250 nt minimal contig length was retained as it corresponds to an encoded 83 amino acids sequence that was felt sufficient for many conserved protein domain searches which are often used in virome analysis or annotation [47].

Effects of community complexity on virome description performance

We evaluated how, for a given sequencing depth, community complexity affects virome description performance following contigs assembly. For this, all datasets were normalized at a 120K read depth. Similar to the initial analysis using reads mapping, the number of detected viruses was reduced as community complexity increased. Again, dsRNA outperformed VANA at all complexity levels, though the difference in performance remained limited for low to medium community complexities (Supplementary Figure S3). VANA performance degradation was however more drastic at high community complexity, dropping from 44 RNA viruses and four DNA viruses detected for communities of five viruses (66.7% of total viruses) to 11 RNA viruses and one DNA virus detected (16.7%) for the 60-viruses community. The corresponding values for dsRNA were 53 (80.3%) and 26 RNA viruses (39.4%), with no DNA virus detected (Supplementary Figure S3). Remarkably, RNASeq turned out to be the least affected, with respectively 57/71 (80.3%, 5-viruses communities) and 34/71 viruses (47.9%, 60-viruses community) detected. These results indicate that even for limited complexity communities involving only five viruses, read numbers significantly higher than 120K are needed by the various techniques to achieve a 100% detection performance with a wide range of viruses.

If trying to compensate increased virome complexity by a parallel increase in sequencing depth, a negative impact of complexity is still seen but is much less severe. For example, for the most complex community of 60 viruses at a 1.44M depth (12*120K reads), VANA detected 23 RNA viruses and 2 DNA viruses (compared to 44 RNA viruses and four DNA viruses when analysing individually the 12 pools of five viruses at 120K reads depth), which corresponds to a reduction in performance of 47.9%. For its part, dsRNA detected 42 RNA viruses (no DNA virus), to be compared with 53 viruses when individually analyzing the 12 pools of five viruses, corresponding to a reduction in performance of 20.7% (Supplementary Figure S4). The corresponding value for RNASeq was 55 viruses detected, corresponding to a performance

386 equivalent to the analysis of the 12 communities of five viruses. The loss in performance
387 resulting from high community complexity is therefore only significant for the dsRNA and
388 VANA approaches, and strongest in the case of VANA.

389 **Impact of sequencing depth on *de novo* assembly**

390 The 60-viruses community was used to investigate the influence of sequencing depth on *de*
391 *novo* assembly performance itself. The VANA and dsRNA datasets were therefore resampled
392 at different depths (100K, 300K, 1M, 3M and 10M reads, five random resampling at each
393 depth), assembled and the obtained contigs mapped against the viral reference genomes to
394 determine the average assembly parameters and viral contigs parameters. The results are shown
395 in Supplementary Table S2 and, for viral contigs alone, in Table 3.

396 As expected, all assembly parameters (number of contigs, average contig length, N50, maximal
397 contig length) increased with sequencing depth (Supplementary Table S2). The same tends to
398 be true for viral contigs (number and length, Table 3), while the proportion of viral contigs
399 tended to diminish as sequencing depth increased, likely reflecting increased probability of
400 assembly of non-viral reads (Supplementary Table S2). Although at the lowest 100K reads
401 sequencing depth few assembly parameters were found to be statistically different, both the
402 total number of assembled contigs and the number of viral contigs were found to be highly
403 statistically different, with dsRNA yielding about 3-fold more contigs and 3-fold more viral
404 contigs than VANA (Table 3 and Supplementary Table S2). This trend was observed at all
405 sequencing depth, with 1.3 to 1.8-fold more viral contigs observed for dsRNA.

406 At other sequencing depths, differences between the VANA and dsRNA assemblies proved
407 systematically highly significant, with dsRNA consistently yielding more numerous and longer
408 contigs as well as more numerous and longer viral contigs. On the other hand, the proportion

409 of viral contigs was found consistently higher in assemblies of the VANA datasets
410 (Supplementary Table S2).

411 It should be noted that the better assembly performance of dsRNA is independent of minimal
412 contig length (Table 4). In particular, using the most complex community and 10 million reads
413 datasets, the higher performance of dsRNA over VANA was observed for all assembly
414 parameters (number of contigs, average length, N50, maximum length) and for both viral
415 contigs parameters (number and average length) at all minimal contigs length (from 125 to 1000
416 nt) with a single exception, the number of viral contigs >125 nt long (1,852 for VANA vs 1,672
417 for dsRNA) (Table 4). At all other minimal contig length, VANA showed from 19.2% (contigs
418 ≥175 nt) to 50.7% (>1 kb contigs) fewer viral contigs than dsRNA and these contigs were 23-
419 33% shorter on average than the dsRNA ones (Table 4).

420 As compared to VANA and dsRNA assemblies, RNASeq assemblies generated more viral
421 contigs at low sequencing depth (ca. 10-30% more than dsRNA for depth of 100K to 1M reads)
422 but ca. 15% fewer viral contigs at the 10M depth. On the other hand, a striking difference in the
423 length of viral contigs was also observed, with RNASeq contigs increasing from an average of
424 1kb (100K depth, 34% longer than dsRNA contigs on average) to 2.1kb (10M depth, 89%
425 longer than for dsRNA).

426 **Impact of sequencing depth on virus identification performance**

427 We proceeded to evaluate the performance of VANA and dsRNA in identifying the expected
428 viruses or viral molecules as affected by sequencing depth. The contigs obtained for the various
429 datasets resampled at different depths (five resampling per sequencing depth) were mapped on
430 individual reference sequences. This allowed to evaluate both the proportion of detected viruses
431 and the coverage of the detected viral molecules, together with their standard deviation
432 (Supplementary Figure S5). Once again, at all sequencing depths and for both parameters,

433 dsRNA outperformed VANA for RNA viruses, while VANA outperformed dsRNA for DNA
434 viruses. In all cases, average coverage of detected segments of RNA viruses showed a high
435 standard deviation but dsRNA contigs covered 9% to 22% more of the detected molecules than
436 VANA contigs.

437 Similarly, and as expected from single reads mapping data, dsRNA outperformed VANA for
438 the identification of RNA viruses present in the most complex, 60-viruses community. For
439 VANA, performance ranged from 17.7% of RNA viruses identified at the 100K reads depth to
440 60.3% at the 10 million reads depth. The corresponding values for dsRNA are respectively
441 35.2% and 89.7% and those for RNASeq respectively 46.2% and 90.8%. The performance of
442 RNASeq therefore appears to be nearly identical to that of dsRNA for RNA viruses, and
443 superior for DNA viruses with 5/6 viruses detected for the 3M and 10M reads depth.

444 A plot of the observed proportion of detected RNA viruses over a logarithmic scale of the
445 sequencing effort is shown in Figure 4. It shows a remarkable pattern with linear regression r^2
446 coefficients of 0.97-0.99, suggesting a very strong and monotonous relationship between
447 sequencing depth and the proportion of the viruses present in the community that are
448 represented by at least one assembled contig. An extension of that trend would suggest that a
449 depth of about 30 million reads would be needed for the dsRNA approach to recover at least
450 one contig for each of the 66 RNA viruses present in the synthetic community, while in excess
451 of 1 billion reads would be needed to achieve a comparable performance using VANA. If taking
452 into account also DNA viruses to calculate a proportion of detected viruses, similar linear
453 relationships are still observed, but the performance of the dsRNA approach is slightly degraded
454 as expected from its poor ability to detect DNA viruses (Figure 4). Analyzed in a similar
455 fashion, the RNASeq data showed the same linear relationship, although with a slightly lower
456 r^2 value of 93.7% and a predicted detection of all 71 viruses and satellites with 16-17M reads.

Due to a more limited number of reads available for virus communities up to the 20-viruses pools, a similar evaluation could not be as extensively performed for these lower complexity communities. However, an analysis at three sequencing depths (100K reads, 300K reads, 875K reads) of the 20-viruses communities data provided comparable results with r^2 correlation coefficients of 0.95-0.98, suggesting that the linear correlation between the percentage of viruses recovered and the log of the sequencing depth is independent of the complexity of the analyzed community (result not shown).

An analysis performed at the level of individual viral genomic molecules (115 viral molecules) allows to evaluate the performance of the two methods using the most complex, 60-viruses pool, for groups of viruses with different genome types. The numbers of viral molecules are however small for RNA satellites, dsRNA viruses and dsDNA viruses. The results, using a 10 million reads sequencing depth, are summarized in Table 5. Considering individual molecules, VANA had at least one contig for only 50% of the viral molecules present in the most complex synthetic community, to be compared with a 76.5% value for dsRNA. But while the VANA performance was at an intermediate level for all virus groups analyzed, dsRNA showed good performance for +ssRNA viruses (89.5% of molecules), RNA satellites (100%) and dsRNA viruses (100%). The dsRNA performance was however poor for DNA viruses, as expected, but also for -ssRNA viruses (41.7% of detected molecules only).

DISCUSSION

While synthetic communities have been widely used to benchmark metagenomic processes targeting bacteria and fungi, methodological benchmarking approaches in virome studies are still limited and largely confined to clinical settings [38, 48-49] and, to some extent, to environmental virome studies [50-51]. Such approaches are today largely lacking in plant virology. Here we used well authenticated and sequence characterized plant virus isolates from

481 a public bioresource center (Leibniz-Institute DSMZ) that allowed for the simple construction
482 of synthetic viral communities of varying complexity. Although some of the viruses were
483 detected by only very low read numbers, no virus was fully absent from all generated datasets,
484 validating the approach and the samples used. The fact that some viruses were identified only
485 by low read numbers could have a variety of reasons, such as low virus titer in some samples,
486 competition with other viruses for reads representation in the assembled communities, or
487 difficulties in extracting viral nucleic acids from some plant species. In addition, the fact that
488 freeze-dried plant material was used in this study may have had a negative impact on results
489 and the analysis of fresh plant tissues might have provided superior results. In this respect, it
490 should be noted that the two viruses present as infected banana samples, banana streak OL virus
491 (BSOLV) and banana bunchy top virus (BBTV), were only detected by very low read numbers
492 using both VANA and dsRNA, despite the fact that these techniques have successfully been
493 used in the past to analyze banana samples [52-53]. The RNASeq data on the same viral isolates
494 shows about 0.9% of viral reads BSOLV but BBTV was the individual sample with the fewest
495 reads by far in the RNASeq analysis, suggesting a low viral concentration in that particular
496 sample.

497 A total of 11 additional viruses or viral agents were identified in the constructed communities.
498 In most cases, these correspond to satellites that had not been specifically indexed in the viral
499 isolates used or of viruses latently infecting propagation hosts, such as Hordeum vulgare
500 endornavirus, which is present in many barley varieties, or Chenopodium quinoa mitovirus.

501 The communities assembled cover all known plant virus genome types, 21 viral families (plus
502 satellites and one virus unassigned in a family) and a total of 61 genera [plus four viruses not
503 currently assigned to a genus and three satellites]. It is thus probably to date the largest scale
504 effort to build synthetic viral communities and use them for the benchmarking of phytovirome
505 analysis approaches. In some benchmarking studies, the nucleic acid proportions of the

individual viruses involved in the virus community were quantified prior to extraction [36, 39]. The fact that no special effort was made here to normalize or measure the concentration of the different viruses is a limitation for some comparisons. On the other hand, the samples used involved different propagation hosts and actual virus titers in those hosts, so that the communities assembled reflect actual samples from plant virome studies. The results obtained indicate that a range of parameters impact the completeness of the virome description achieved. Not surprisingly, such parameters include (i) sequencing depth, (ii) community complexity, (iii) use of *de novo* assembled contigs *vs* use of unassembled reads and (iv) minimal contig length.

The key objective of this work was to compare the performance of the VANA and dsRNA approaches, which are the two techniques most widely used in ecology-oriented viral metagenomics experiments involving the analysis of complex pools of plants. The results provided here for RNASeq following ribodepletion should be considered with caution, since they are not fully comparable with the VANA or dsRNA data. Indeed, the RNASeq datasets for the various communities were assembled *in silico*, from data obtained by single-isolate sequencing. This means that any interactions between plant samples or competition between viruses for representation in the datasets were eliminated, contrary to the situation with the VANA and dsRNA experiments. Given that RNASeq is considered an unbiased approach (hence its use for transcriptome analysis), this should not be a problem but the existence of unforeseen effects affecting the results cannot be completely ruled out. As compared to dsRNA and VANA, the results obtained for RNASeq using the *in silico* assembled communities show: (i) a much lower imbalance in the representation of the various viruses (3 logs variation as opposed to 5-6 logs), (ii) on average significantly longer viral contigs, irrespective of sequencing depth and (iii) an overall excellent performance with 90% of the viruses identified at 10M reads depth for the most complex, 60-viruses community. This last result favourably compares with the dsRNA performance for all viral categories with the exception of viruses

with dsRNA genomes (Table 5). This performance comes as a surprise given the absence of enrichment (besides ribodepletion) in RNASeq. However, the relatively narrow range of variation in the proportion of viral reads for different viruses, possibly implying reduced competition for representation between viruses, and the even distribution of RNASeq reads along viral genomes, possibly favouring a more efficient genome assembly, could have contributed to the RNASeq performance. In any case, these results surprisingly suggest that RNASeq could have a very good potential for the analysis of complex viral communities and clearly call for direct benchmarking efforts using RNASeq and complex synthetic or natural communities in order to unambiguously validate this potential.

As previously reported using natural communities (Ma *et al.*, 2019), the dsRNA approach provided in all comparisons a more complete description of the RNA virome than the VANA approach but performed very poorly with DNA viruses. However, the differential with VANA is more limited for the less complex communities of five or 10 viruses. According to our own experience, this level of complexity is most often seen when analyzing single plants or pools of 5-20 plants of the same species, with vegetatively propagated plants tending to have more complex viromes. Higher complexity levels are usually encountered when analyzing larger pools composed of plants belonging to different species. The dsRNA approach is therefore recommended whenever analysing complex viromes or when an emphasis on RNA viruses is of importance, in particular since dsRNA allows comparable levels of completeness with a lower sequencing effort. On the other hand, for viromes of low to medium complexity, the results reported here show VANA to be a reasonable alternative. For example, at 480K reads depth, VANA detected 57.4% of all viruses for the 20-viruses communities as compared to 61.8% for dsRNA (result not shown, see also Supplementary Figure 4 for the compared rates of detection of RNA viruses only). VANA should of course be the preferred choice if analysis of DNA viruses is of importance. The reason for the better performance of the dsRNA approach

556 for high complexity viromes is not fully clear but might result from a lower level of competition
557 between viral nucleic acid molecules for representation in complex pools, resulting in a
558 somewhat less imbalanced distribution of read numbers between viruses (Figure 2). Different
559 human microbiome studies have shown that different steps of RNA/DNA extraction such as
560 homogenization, centrifugation, filtration and chloroform treatment, can have a major impact
561 on the quantitative and qualitative composition of identified viral communities, skewing viral
562 metagenome assemblies [37-38, 54]. Another critical step is library preparation, which often
563 involves a random amplification PCR to increase virus genetic material and to add linkers,
564 allowing samples multiplexing during HTS sequencing and thus reducing sequencing costs.
565 The amplification step may alter the relative abundance of viruses and can lead to uneven
566 coverage if random primers do not anneal randomly on viral genomes. Indeed, in the case of
567 faba bean necrotic stunt virus, the relative frequencies of the different genome segments
568 determined by qPCR was significantly different before and after a rolling circle amplification
569 step used prior to HTS sequencing [55]. Furthermore, different library preparation techniques
570 have been found to require different sequencing depths to achieve the same genome coverage
571 [56]. Regardless of the experiment, it is advisable to develop an estimate of the sequencing
572 depth needed, so as to be able to answer the biological question at hand while avoiding
573 excessive sequencing costs. Here we identified a very robust correlation between the percentage
574 of viruses identified in complex communities and the log of the sequencing depth. This is an
575 interesting result, since it allows to gauge the sequencing effort needed for a particular level of
576 virome description or, conversely, to gauge the extent of virome description that can be
577 expected from a particular sequencing depth. Besides metagenomic studies, this finding might
578 have practical implications for diagnostics since many plants, in particular vegetatively
579 propagated ones, frequently display complex mixed infections involving a range of viruses.

Virus detection in metagenomic studies is constrained by the degree of complexity of the virus communities analyzed. Our results suggest that the detection efficiency of either mapping of unassembled reads or analysis of *de novo* assembled contigs were affected by community complexity with a general trend of detecting a lower proportion of viruses in more complex communities. However, the read mapping strategy was more efficient at all complexities (Figure 1B and Supplementary Figure S3), confirming results obtained through performance testing of sequence analysis strategies [57]. This may be due to the complexity of *de novo* assembly of complex communities, linked with insufficient coverage or uneven coverage of low abundance viruses within such communities. Correspondingly, we observed a lower virus detection rate when using longer minimal contig sizes in the *de novo* assembly, which again might be attributed to difficulties in assembling reads from more complex communities for example when coexisting viruses share highly similar regions in their genomes, leading to higher fragmentation and reduced contig sizes [58].

Lastly, it has been reported that the quality and completeness of virome description is also affected by the bioinformatic analysis used [58-61]. The normalized 10M reads datasets generated in the present study with the 60-viruses community, which are available at https://doi.org/10.57745/T4UYPC, together with the community composition and the complete or near complete reference genomic sequences used here should prove very useful tools to benchmark virome characterization pipelines.

Acknowledgements: The authors wish to thank the INRAE GetPlaGe Platform (GenoToul, Toulouse, France) for Illumina sequencing.

Funding: This study was funded by the European Union through a Horizon 2020 Marie Skłodowska-Curie Actions Innovative Training Network (H2020 MSCA- 60 ITN) project "INEXTVIR" (GA 813542). The individual sequencing of some of the isolates used in this study was performed as part of the EVA-Global project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871029. OM is recipient of a PhD fellowship from CIRAD and the ANR (Phytovirus project number: ANR-19-CE35-0008-02).

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC (2019). Expanding the RNA Virosphere by Unbiased Metagenomics. Annual Review of Virology, *6*, 119–139. https://doi.org/10.1146/annurev-virology-092818-015851

[2] Jian H, Yi Y, Wang J, Hao Y, Zhang M, Wang S, Meng C, Zhang Y, Jing H, Wang Y, Xiao X (2021) Diversity and distribution of viruses inhabiting the deepest ocean on Earth. ISME J. 15, 3094-3110. https://doi.org/10.1038/s41396-021-00994-y

[3] Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A (2019) Evolution and ecology of plant viruses. Nature Reviews Microbiology 17, 632-644. https://doi.10.1038/s41579-019-0232-3

[4] Maclot F, Candresse T, Filloux D, Malmstrom CM, Roumagnac P, van der Vlugt R, Massart S (2020) Illuminating an ecological blackbox: using high throughput sequencing to characterize the plant virome across scales. Front. Microbiol. 11, 578064. https://doi.org/10.3389/fmicb.2020.578064

[5] Roux S, Matthijnssens J, Dutilh BE (2019) Metagenomics in Virology. Reference Module in Life Sciences. https://doi.org/10.1016/B978-0-12-809633-8.20957-6:B978-0-12-809633-8.20957-6

[6] Greninger AL (2018) A decade of RNA virus metagenomics is (not) enough. Virus Research 244, 218-229. https://doi.org/10.1016/j.virusres.2017.10.014

[7] Moubset O, François S, Maclot F, Palanga E, Julian C, Claude L, Fernandez E, Rott P, Daugrois JH, Antoine-Lorquin A, Bernardo P, Blouin AG, Temple C, Kraberger S, Fontenele RS, Harkins GW, Ma Y, Marais A, Candresse T, Chéhida SB, Lefeuvre P, Lett JM, Varsani A, Massart S, Ogliastro M, Martin DP, Filloux D, Roumagnac P (2022) Virion-associated nucleic acid-based metagenomics: a decade of advances in molecular characterization of plant viruses. Phytopathology, 112, 2253-2272. https://doi.org/10.1094/PHYTO-03-22-0096-RVW

[8] Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, Chen IM, Ivanova N, Allen LZ, Paez-Espino D, Bryant DA, Bhaya D, Consortium RVD, Krupovic M, Dolja VV, Kyrpides NC, Koonin EV, Gophna U (2022) A five-fold expansion of the global RNA virome reveals multiple new clades of RNA bacteriophages. bioRxiv https://doi.org/10.1016/j.cell.2022.08.023

[9] Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ (2018) The evolutionary history of vertebrate RNA viruses. Nature, 556, 197-202. https://doi.org/10.1038/s41586-018-0012-7

[10] Wren JD, Roossinck MJ, Nelson RS, S2cheets K, Palmer MW, Melcher U (2006) Plant virus biodiversity and ecology. PLoS Biol. 4, e80. https://doi.org/10.1371/journal.pbio.0040080

[11] Kobayashi K, Atsumi G, Iwadate Y, Tomita R, Chiba K, Akasaka S, Nishihara M, Takahashi H, Yamaoka N, Nishiguchi M, Sekine K (2013) Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome. J. Gen. Plant Pathol. 79, 56-63. https://doi.org/10.1007/s10327-012-0423-5

[12] Schönegger D, Marais A, Faure C, Candresse T (2022) A new flavi-like virus identified in populations of wild carrots. Arch Virol. 167, 2407-2409. https://doi: 10.1007/s00705-022-05544-1

656 [13] Roossinck MJ, Martin DP, Roumagnac P (2015) Plant virus metagenomics: advances in
657 virus discovery. Phytopathology, 105, 716-727. https://doi.org/10.1094/PHYTO-12-14-
658 0356-RVW

659 [14] Maree HJ, Fox A, Al Rwahnih M, Boonham N, Candresse T (2018) Application of HTS
660 for routine plant virus diagnostics: state of the art and challenges. Front. Plant Sci. 9,
661 1082. https://doi.org/10.3389/fpls.2018.01082

662 [15] Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P (2004)
663 Emerging infectious diseases of plants: pathogen pollution, climate change and
664 agrotechnology drivers. Trends Ecol. Evol. 19, 535-544.
665 https://doi.org/10.1016/j.tree.2004.07.021

666 [16] Candresse T, Marais A, Faure C, Gentit P (2013) Association of Little cherry virus 1 with
667 the Shirofugen stunt disease and characterization of the genome of a divergent LChV1
668 isolate. Phytopathology 103, 293-298. https://doi.org/10.1094/PHYTO-10-12-0275-R

669 [17] Moreno AB, López-Moya JJ (2020) When viruses play team sports: mixed infections in
670 plants. Phytopathology, 110, 29-48. https://doi.org/10.1094/PHYTO-07-19-0250-FI

671 [18] Malmstrom CM, Melcher U, Bosque-Pérez NA (2011) The expanding field of plant virus
672 ecology: historical foundations, knowledge gaps, and research directions. Virus Res. 159,
673 84-94. https://doi.org/10.1016/j.virusres.2011.05.010

674 [19] Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, Hartnady
675 P, Rebelo T A, Cousins SR, Mesleard F, Cohez D, Yavercovski N, Varsani A, Harkins
676 GW, Peterschmitt M, Malmstrom CM, Martin DP, Roumagnac P (2018)
677 Geometagenomics illuminates the impact of agriculture on the distribution and
678 prevalence of plant viruses at the ecosystem scale. ISME J. 12, 173-184.
679 https://doi.org/10.1038/ismej.2017.155

680 [20] Ma Y, Marais A, Lefebvre M, Faure C, Candresse T (2020) Metagenomic analysis of
681 virome cross-talk between cultivated Solanum lycopersicum and wild Solanum nigrum.
682 Virology, 540, 38-44. https://doi.org/10.1016/j.virol.2019.11.009

683 [21] Ma Y, Fort T, Marais A, Lefebvre M, Theil S, Vacher C, Candresse T (2021) Leaf-
684 associated fungal and viral communities of wild plant populations differ between

685 cultivated and natural ecosystems. Plant Environ. Interact. 2, 87-99.
686 https://doi.org/10.1002/pei3.10043

687 [22] Susi H, Laine AL (2021) Agricultural land use disrupts biodiversity mediation of virus
688 infections in wild plant populations. New Phytol. 230, 2447-2458.
689 https://doi.org/10.1111/nph.17156

690 [23] Maachi A, Donaire L, Hernando Y, Aranda MA (2022) Genetic differentiation and
691 migration fluxes of viruses from melon crops and crop edge weeds. J. Virol. 96, e00421-
692 22. https://doi.org/10.1128/jvi.00421-22

693 [24] Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009)
694 Complete viral genome sequence and discovery of novel viruses by deep sequencing of
695 small RNAs: A generic method for diagnosis, discovery and sequencing of viruses.
696 Virology, 388, 1-7. https://doi.org/10.1016/j.virol.2009.03.024

697 [25] Kashif M, Pietila S, Artola K, Jones RAC, Tugume AK, Mäkinen V, Valkonen JPT
698 (2012) Detection of viruses in sweet potato from Honduras and Guatemala augmented by
699 deep-sequencing of small RNAs. Plant Dis. 96, 1430-1437.
700 https://doi.org/10.1094/PDIS-03-12-0268-RE

701 [26] Thapa V, McGlinn DJ, Melcher U, Palmer MW, Roossinck MJ (2015) Determinants of
702 taxonomic composition of plant viruses at the Nature Conservancy's Tallgrass Prairie
703 Preserve, Oklahoma. Virus Evol. 1, 1-8. https://doi.org/10.1093/ve/vev007

704 [27] Villamor DEV, Ho T, Al Rwahnih M, Martin RR, Tzanetakis IE (2019) High throughput
705 sequencing for plant virus detection and discovery. Phytopathology, 109, 716-725.
706 https://doi.org/ 10.1094/PHYTO-07-18-0257-RVW

707 [28] Kutnjak D, Tamisier L, Adams I, Boonham N, Candresse T, Chiumenti M, De Jonghe K,
708 Kreuze JF, Lefebvre M, Silva G, Malapi-Wight M, Margaria P, Mavrič Pleško I, McGreig
709 S, Miozzi L, Remenant B, Reynard JS, Rollin J, Rott M, Schumpp O, Massart S,
710 Haegeman A (2021) A primer on the analysis of high-throughput sequencing data for
711 detection of plant viruses. Microorganisms, 9, 841.
712 https://doi.org/10.3390/microorganisms9040841

713 [29] Pecman A, Kutnjak D, Gutiérrez-Aguirre I, Adams I, Fox A, Boonham N, Ravnikar M
714 (2017) Next Generation Sequencing for Detection and Discovery of Plant Viruses and
715 Viroids: Comparison of Two Approaches. Front Microbiol. 8, 1998.
716 https://doi.org/10.3389/fmicb.2017.01998

717 [30] Gaafar YZA, Ziebell H (2020) Comparative study on three viral enrichment approaches
718 based on RNA extraction for plant virus/viroid detection using high-throughput
719 sequencing. PLoS ONE, 15, 1-17. https://doi.org/10.1371/journal.pone.0237951

720 [31] Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois JH,
721 Fernandez E, Martin DP, Varsani A, Roumagnac P (2014) Appearances can be deceptive:
722 revealing a hidden viral infection with deep sequencing in a plant quarantine context.
723 PLoS ONE, 9, e102945. https://doi.org/10.1371/journal.pone.0102945

724 [32] Ma Y, Marais A, Lefebvre M, Theil S, Svanella-Dumas L, Faure C, Candresse T (2019)
725 Phytovirome analysis of wild plant populations: comparison of double-stranded RNA and
726 virion-associated nucleic acids metagenomic approaches. J. Virol. 94, e01462-19.
727 https://doi.org/10.1128/jvi.01462-19

728 [33] Massart S, Adams I, Al Rwahnih M, Baeyen S, Bilodeau GJ, Blouin A, BoonhamN,
729 Candresse T, Chandellier A, De Jonghe K, Fox A, Gaafar YZA, Gentit P, Haegeman A,
730 Ho W, Hurtado-Gonzales O, Jonkers W, Kreuze J, Kutjnak D, Landa BB, Liu M, Maclot
731 F, Malapi-Wight M, Maree HJ, Martoni F, Mehle N, Minafra A, Mollov D, Moreira AG,
732 Nakhla M, Petter F, Piper AM, Ponchart JP, Rae R, Remenant B, Rivera Y, Rodoni B,
733 Botermans M, Roenhorst JW, Rollin J, Saldarelli P, Santala J, Souza-Richards R, Spadaro
734 D, Studholme DJ, Sultmanis S, van der Vlugt R, Tamisier L, Trontin C, Vazquez-Iglesias
735 I, Vicente CSL, van de Vossenberg BTLH, Westenberg M, Wetzel T, Ziebell H, Lebas
736 BSM (2022) Guidelines for the reliable use of high throughput sequencing technologies
737 to detect plant pathogens and pests. Peer Comm. J. 2, e62.
738 https://doi.org/10.24072/pcjournal.181

739 [34] Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of
740 a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence
741 data on the MiSeq illumina sequencing platform. Appl. Environ. Microbiol. 79, 5112-
742 5120. https://doi.org/10.1128/AEM.01043-13

743 [35] Egan CP, Rummel A, Kokkoris V, Klironomos J, Lekberg Y, Hart M (2018) Using mock
744 communities of arbuscular mycorrhizal fungi to evaluate fidelity associated with Illumina
745 sequencing. Fungal Ecol. 33, 52-64. https://doi.org/10.1016/j.funeco.2018.01.004

746 [36] Sevim V, Lee J, Egan R, Clum A, Hundley H, Lee J, Everroad RC, Detweiler AM, Bebout
747 BM, Pett-Ridge J, Göker M, Murray AE, Lindemann SR, Klenk HP, O'Malley R, Zane
748 M, Cheng JF, Copeland A, Daum C, Woyke T (2019) Shotgun metagenome data of a
749 defined mock community using Oxford Nanopore, PacBio and Illumina technologies.
750 Sci. Data, 6, 285. https://doi.org/10.1038/s41597-019-0287-z

751 [37] Conceição-Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L, Deboutte W, Yinda CK,
752 Lavigne R, Maes P, Ranst M Van, Heylen E, Matthijnssens J (2015) Modular approach
753 to customise sample preparation procedures for viral metagenomics: A reproducible
754 protocol for virome analysis. Sci. Rep. 5, 1-14. https://doi.org/10.1038/srep16532

755 [38] Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A (2018)
756 Evaluation of bias induced by viral enrichment and random amplification protocols in
757 metagenomic surveys of saliva DNA viruses. Microbiome, 6,1-18.
758 https://doi.org/10.1186/s40168-018-0507-3

759 [39] Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman
760 ML, Breitbart M, Sullivan MB (2016) Towards quantitative viromics for both double-
761 stranded and single-stranded DNA viruses. Peer J. 4, e2777.
762 https://doi.org/10.7717/peerj.2777

763 [40] Gil P, Dupuy V, Koual R, Exbrayat A, Loire E, Fall AG, Gimonneau G, Biteye B, Talla
764 Seck M, Rakotoarivony I, Marie A, Frances B, Lambert G, Reveillaud J, Balenghien T,
765 Garros C, Albina E, Eloit M, Gutierrez S (2021) A library preparation optimized for
766 metagenomics of RNA viruses. Mol. Ecol. Res. 21, 1788-1807.
767 https://doi.org/10.1111/1755-0998.13378

768 [41] Marais A, Faure C, Bergey B, Candresse T (2018) Viral double-stranded RNAs (dsRNAs)
769 from Plants: Alternative nucleic acid substrates for high-throughput sequencing. In: Viral
770 Metagenomics: Methods in Molecular Biology, Pantaleo V, Chiumenti M (Eds.),
771 Humana Press, New York (USA), pp 45-53.

772 [42] François S, Filloux D, Fernandez E, Ogliastro M, Roumagnac P (2018) Viral
773 metagenomics approaches for high-resolution screening of multiplexed arthropod and
774 plant viral communities. In: Viral metagenomics: methods and protocols, Pantaleo V,
775 Chiumenti M (Eds), Humana Press, New York (USA), pp 77-95.

776 [43] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment
777 search tool. J Mol Biol. 215 403-410. https://doi: 10.1016/S0022-2836(05)80360-2

778 [44] Illumina (2017) Effects of index misassignment on multiplexing and downstream
779 analysis. https://www.illumina.com/content/dam/illumina-
780 marketing/documents/products/ whitepapers/index-hopping-white-paper-770-2017-
781 004.pdf

782 [45] van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K (2019) Index hopping on
783 the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol.
784 Resour. 20, 1171-1181 https://doi.org/10.1111/1755-0998.13009

785 [46] François S, Filloux D, Frayssinet M, Roumagnac P, Martin DP, Ogliastro M, Froissart R
786 (2018) Increase in taxonomic assignment efficiency of viral reads in metagenomic
787 studies. Virus Res. 244, 230-234. https://doi.org/10.1016/j.virusres.2017.11.011

788 [47] Lefebvre M, Theil S, Ma Y, Candresse T (2019) The VirAnnot pipeline: A resource for
789 automated viral diversity estimation and operational taxonomy units assignation for
790 virome sequencing data. Phytobiomes J. 3, 256-259. https://doi.org/10.1094/PBIOMES-
791 07-19-0037-A

792 [48] Ajami NJ, Wong MC, Ross MC, Lloyd RE, Petrosino F (2018). Maximal viral
793 information recovery from sequence data using VirMAP. Nat. Comm. 9, 3205.
794 https://doi.org/10.1038/s41467-018-05658-8

795 [49] Santiago-Rodriguez TM, Hollister EB (2020) Potential applications of human viral
796 metagenomics and reference materials: considerations for current and future viruses.
797 Appl. Environ. Microbiol. 86, e01794-20. https://doi.org/10.1128/AEM.01794-20

798 [50] Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko
799 N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S,
800 Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Tara Oceans Coordinators, Bork

P, Acinas SG, Wincker P, Sullivan MB (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature, 537, 689-693. https://doi.org/10.1038/nature19366

[51] Zablocki O, Michelsen M, Burris M, Solonenko N, Warwick-Dugdale J, Ghosh R, Pett-Ridge J, Sullivan MB, Temperton B (2021) VirION2: A shortand long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. Peer J. 9, 1-23. https://doi.org/10.7717/peerj.11088

[52] Filloux D, Dallot S, Delaunay A, Galzi S, Jacquot E, Roumagnac P (2015) Metagenomics approaches based on virion-associated nucleic acids (VANA): An innovative tool for assessing without a priori viral diversity of plants. Meth. Molec. Biol. 1302, 249-257. https://doi.org/10.1007/978-1-4939-2620-6_18

[53] Teycheney PY, Bandou E, Gomez RM, LangeD, Pavis C, Umber M, Acina Manbole IN, Bonheur L, Daugrois JH, Fernandez E, Filloux D, Julian C, Roumagnac P, Grisoni M, Pierret A, Rubington M, Candresse T, Contreras S, Faure C, Marais A, Theil S, Da Câmara MA, Mendonça D, Pinheiro de Carvalho M (2015). Viral treasure hunt in European outermost territories: how metagenomics boosts the discovery of novel viral species in tropical and sub-tropical crops germplasm. https://agritrop.cirad.fr/575812/

[54] Kleiner M, Hooper LV, Duerkop BA (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics, 16, 7. https://doi.org/10.1186/s12864-014-1207-4

[55] Gallet R, Fabre F, Michalakis Y, Blanc S (2017) The number of target molecules of the amplification step limits accuracy and sensitivity in ultradeep-sequencing viral population studies. J Virol. 91, e00561-17. https://doi.org/10.1128/JVI.00561-17

[56] Visser M, Bester R, Burger JT, Maree HJ (2016) Next-generation sequencing for virus detection: covering all the bases. Virol J. 13, 85. https://doi.org/10.1186/s12985-016-0539-x

[57] Massart S, Chiumenti M, De Jonghe K, Glover R, Haegeman A, Koloniuk I, Komínek P, Kreuze J, Kutnjak D, Lotos L, Maclot F, Maliogka V, Maree HJ, Olivier T, Olmos A, Pooggin MM, Reynard JS, Ruiz-García AB, Safarova D, Schneeberger PHH, Sela N, Turco S, Vainio EJ, Varallyay E, Verdin E, Westenberg M, Brostaux Y, Candresse T

(2019) Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. Phytopathology, 109, 488-497. https://doi.org/10.1094/PHYTO-02-18-0067-R

[58] Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB (2017) Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. Peer J. 5, e3817. https://doi.org/10.7717/peerj.3817

[59] Breitwieser FP, Lu J, Salzberg SL (2019) A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 20, 1125-1136. https://doi.org/10.1093/bib/bbx120

[60] Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, Candela M (2016) ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics, 17, 1-9. https://doi.org/10.1186/s12864-016-2446-3

[61] Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C (2019) Choice of assembly software has a critical impact on virome characterisation. Microbiome, 7, 1-15. https://doi.org/10.1186/s40168-019-0626-5

Figures

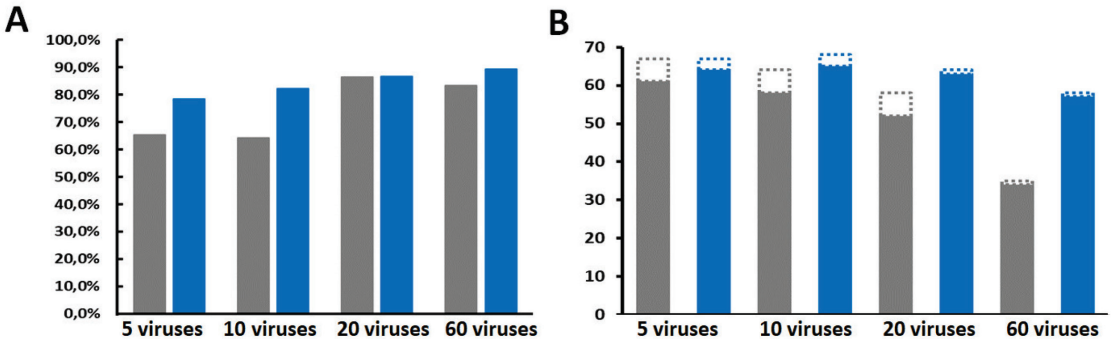


Figure 1. Average proportion of viral reads (DNA and RNA viruses) in VANA (grey) and dsRNA (blue) datasets from viral communities of different complexities (A) and number of viruses detected at an even 120K read depth for communities of different complexities (B). In figure 1B RNA viruses are indicated by solid bars while DNA viruses are indicated by dashed bars.

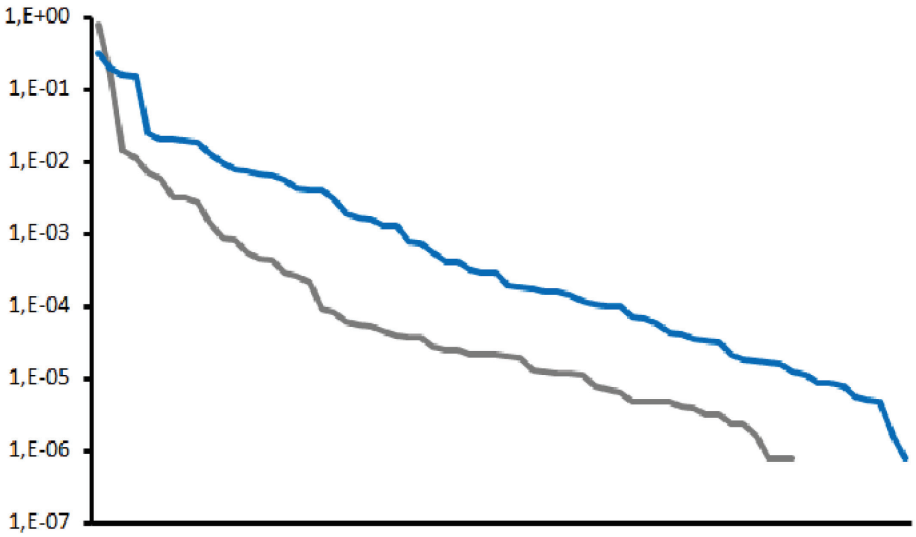


Figure 2: Distribution of percentage of mapped VANA (grey) and dsRNA (blue) reads for each detected virus in the 60-viruses community using a normalized 1.44 million reads sequencing depth. The percentages of mapped reads for each virus are shown on a logarithmic scale, from 1,E+00 (100%) to 1,E-07 (0.000001%)

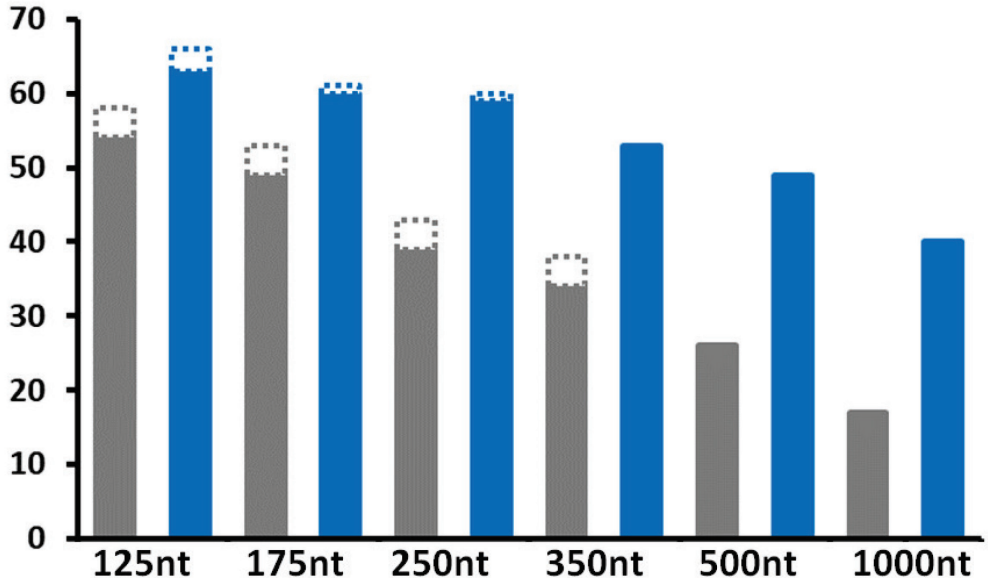
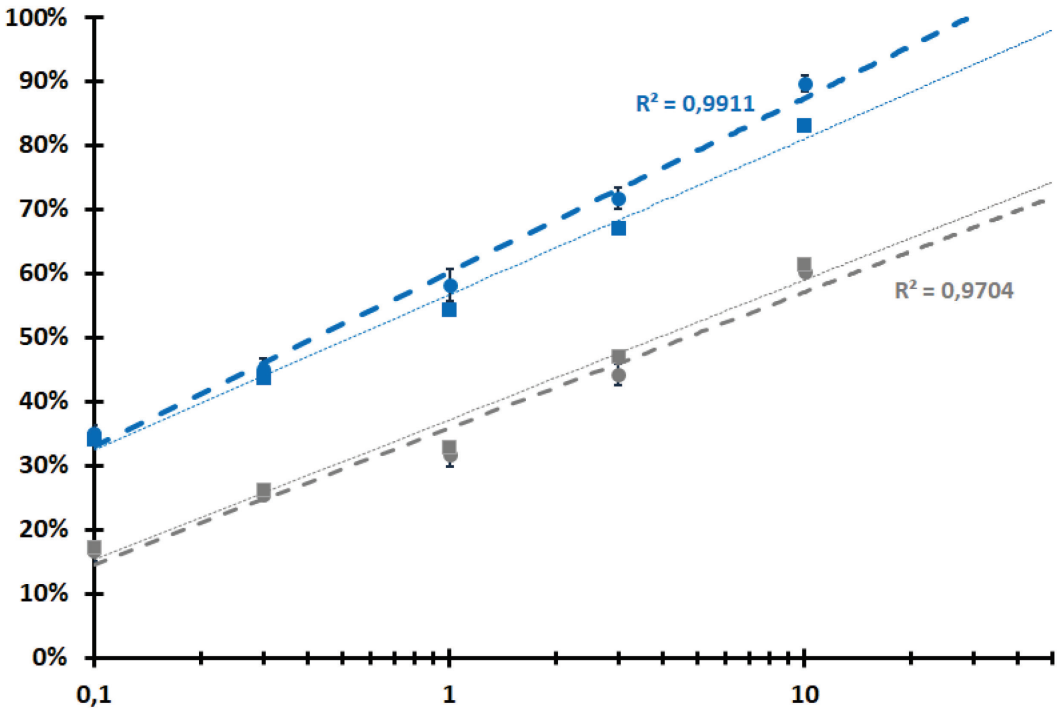


Figure 3. Number of detected viruses using VANA (grey) or dsRNA (blue) in the 60-viruses community (over a total of 69 viruses plus 3 satellites) as a function of minimal contig length at a sequencing depth of 10M reads. RNA viruses are indicated by solid bars while DNA viruses are indicated by dashed bars.



865

866 **Figure 4:** Observed percentages of detected viruses in the 60 viruses community as a function
867 of sequencing depth expressed in million reads per sample and plotted on a logarithmic scale.
868 VANA results are in grey, dsRNA results in blue. Linear regression curves are shown for RNA
869 viruses (round dots, thick lines,) as well as considering both RNA and DNA viruses (square
870 dots, think lines). Linear r^2 coefficients are shown only for the RNA viruses curves.

871

Table 1. Viral isolates used to construct mock viral communities of varying complexity. The taxonomic status of the various viruses is indicated, together with their DSMZ catalogue code, their propagation host and the GenBank accession number(s) of their genomic sequence(s).

Family	Genus	Virus	Acronym	Genome	Code[a]	Host[b]	Sequence accession number(s)
Alphaflexiviridae	*Allexivirus*	Shallot virus X	ShVX	ssRNA(+)	PV-0622	*Chenopodium murale*	MW854280
Alphaflexiviridae	*Potexvirus*	Lettuce virus X	LeVX	ssRNA(+)	PV-0904	*Nicotiana benthamiana*	MW248356
Benyviridae	*Benyvirus*	Beet necrotic yellow vein virus	BNYVV	ssRNA(+)	PV-0467	*Chenopodium quinoa*	OK181765-67; M36896
Betaflexiviridae	*Capillovirus*	Apple stem grooving virus	ASGV	ssRNA(+)	PV-0199	*Chenopodium quinoa*	MW582790
Betaflexiviridae	*Carlavirus*	Poplar mosaic virus	PopMV	ssRNA(+)	PV-0341	*Nicotiana benthamiana*	ON924213
Betaflexiviridae	*Trichovirus*	Apple chlorotic leaf spot virus	ACLSV	ssRNA(+)	PV-0998	*Chenopodium quinoa*	OK340218-19[c]
Betaflexiviridae	*Tepovirus*	Potato virus T	PVT	ssRNA(+)	PV-1145	*Nicotiana hesperis*	MZ405665
Bromoviridae	*Alfamovirus*	Alfalfa mosaic virus	AMV	ssRNA(+)	PV-0779	*Nicotiana tabacum* "Samsun nn"	MZ405653-55
Bromoviridae	*Anulavirus*	Pelargonium zonate spot virus	PZSV	ssRNA(+)	PV-0259	*Nicotiana glutinosa* "24A"	ON398493-95
Bromoviridae	*Bromovirus*	Brome mosaic virus	BMV	ssRNA(+)	PV-0194	*Hordeum vulgare*	MW582787-89
Bromoviridae	*Cucumovirus*	Peanut stunt virus	PSV	ssRNA(+)	PV-0190	*Nicotiana benthamiana*	MW307259-61
Bromoviridae	*Ilarvirus*	Parietaria mottle virus	PMoV	ssRNA(+)	PV-0400	*Chenopodium quinoa*	MZ405646-48
Closteroviridae	*Closterovirus*	Beet yellows virus	BYV	ssRNA(+)	PV-1260	*Beta macrocarpa*	MT815988
Closteroviridae	*Crinivirus*	Tomato chlorosis virus	ToCV	ssRNA(+)	PV-1242	*Solanum lycopersicum*	ON398512-13
Potyviridae	*Bymovirus*	Barley yellow mosaic virus	BaYMV	ssRNA(+)	PV-0634	*Hordeum vulgare*	OL311692-93
Potyviridae	*Ipomovirus*	Cucumber vein yellowing virus	CVYV	ssRNA(+)	PV-0776	*Cucumis sativus*	OK181771
Potyviridae	*Potyvirus*	Bidens mottle virus	BiMoV	ssRNA(+)	PV-0752	*Nicotiana benthamiana*	ON398504
Potyviridae	*Rymovirus*	Agropyron mosaic virus	AgMV	ssRNA(+)	PV-0729	*Triticum aestivum*	OM471970
Potyviridae	*Tritimovirus*	Brome streak mosaic virus	BrSMV	ssRNA(+)	PV-0431	*Hordeum vulgare*	OP357935

Potyviridae	Unassigned	Spartina mottle virus	SpMV	ssRNA(+)	PV-0970	*Spartina* sp.	MN788417
Secoviridae	*Cheravirus*	Arracacha virus B	AVB	ssRNA(+)	PV-0082	*Chenopodium murale*	MW582785-86
Secoviridae	*Comovirus*	Squash mosaic virus	SqMV	ssRNA(+)	PV-0581	*Cucurbita pepo*	ON398498-99
Secoviridae	*Fabavirus*	Broad been wilt virus 1	BBWV-1	ssRNA(+)	PV-0067	*Chenopodium quinoa*	MT663310-11
Secoviridae	*Nepovirus*	Tomato black ring virus	TBRV	ssRNA(+)	PV-0191	*Nicotiana clevelandii*	MW057704-05
Secoviridae	*Sequivirus*	Carrot necrotic dieback virus	CNDV	ssRNA(+)	PV-0976	*Nicotiana benthamiana*	MW080951
Secoviridae	*Stralirivirus*	Strawberry latent ringspot virus	SLRSV	ssRNA(+)	PV-0247	*Chenopodium quinoa*	MZ405640-41
Solemoviridae	*Sobemovirus*	Rice yellow mottle virus	RYMV	ssRNA(+)	PV-0732	*Oryza sativa*	MT701719
Solemoviridae	*Enamovirus*	Pea enation mosaic virus 1	PEMV1	ssRNA (+)	PV-0088	*Pisum sativum*	MW961146
Solemoviridae	*Polerovirus*	Cucurbit aphid-borne yellows virus	CABYV	ssRNA(+)	PV-1017	*Physalis floridana*	MZ202344
Tombusviridae	*Alphacarmovirus*	Calibrachoa mottle virus	CbMV	ssRNA(+)	PV-0611	*Chenopodium quinoa*	OK181769
Tombusviridae	*Alphanecrovirus*	Tobacco necrosis virus A	TNV-A	ssRNA(+)	PV-0186	*Chenopodium quinoa*	MT675968
Tombusviridae	*Aureusvirus*	Johnsongrass chlorotic stripe mosaic virus	JCSMV	ssRNA(+)	PV-0605	*Zea mays*	MT682309
Tombusviridae	*Betacarmovirus*	Turnip crinkle virus	TCV	ssRNA(+)	PV-0293	*Nicotiana benthamiana*	OK181761
Tombusviridae	*Betanecrovirus*	Beet black scorch virus	BBSV	ssRNA(+)	PV-0951	*Chenopodium quinoa*	OK058516
Tombusviridae	*Dianthovirus*	Carnation ringspot virus	CRSV	ssRNA(+)	PV-0097	*Nicotiana clevelandii*	MT682300-01
Tombusviridae	*Gammacarmovirus*	Melon necrotic spot virus	MNSV	ssRNA(+)	PV-0378	*Cucumis sativus*	ON398496
Tombusviridae	*Machlomovirus*	Maize chlorotic mottle virus	MCMV	ssRNA(+)	PV-1087	*Zea mays*	OK181780
Tombusviridae	*Pelarspovirus*	Pelargonium line pattern virus	PLPV	ssRNA(+)	PV-0193	*Chenopodium quinoa*	MW854266
Tombusviridae	*Tombusvirus*	Tomato bushy stunt virus	TBSV	ssRNA(+)	PV-0268	*Nicotiana clevelandii*	MW582792
Tombusviridae	*Umbravirus*	Carrot mottle virus	CMoV	ssRNA(+)	PV-0968	*Nicotiana benthamiana*	OK058520
Tombusviridae	*Umbravirus*	Pea enation mosaic virus 2	PEMV2	ssRNA(+)	PV-0088	*Pisum sativum*	MW961147; MW961148[c]
Tospoviridae	*Orthotospovirus*	Impatiens necrotic spot virus	INSV	ssRNA(+/-)	PV-0280	*Nicotiana benthamiana*	MW582795-97
Tymoviridae	*Tymovirus*	Turnip yellow mosaic virus	TYMV	ssRNA(+)	PV-0299	*Brassica rapa*	ON924209

Virgaviridae	*Furovirus*	Soil-borne wheat mosaic virus	SBWMV	ssRNA(+)	PV-0748	*Triticum aestivum*	MZ405651-52
Virgaviridae	*Hordeivirus*	Barley stripe mosaic virus	BSMV	ssRNA(+)	PV-0330	*Hordeum vulgare*	ON924210-12
Virgaviridae	*Pecluvirus*	Peanut clump virus	PCV	ssRNA(+)	PV-0291	*Nicotiana benthamiana*	MW961156-57
Virgaviridae	*Pomovirus*	Potato mop-top virus	PMTV	ssRNA(+)	PV-0582	*Nicotiana benthamiana*	ON398500-02
Virgaviridae	*Tobamovirus*	Paprika mild mottle virus	PaMMV	ssRNA(+)	PV-0606	*Nicotiana benthamiana*	OK181768
Virgaviridae	*Tobravirus*	Pea early-browning virus	PEBV	ssRNA(+)	PV-0298	*Chenopodium quinoa*	MW854268-69
not assigned	*Idaeovirus*	Raspberry bushy dwarf virus	RBDV	ssRNA(+)	PV-0053	*Chenopodium quinoa*	MW582777-78
Rhabdoviridae	*Cytorhabdovirus*	Lettuce necrotic yellows virus	LNYV	ssRNA(-)	PV-0085	*Nicotiana glutinosa* “24A”	MZ202327
Rhabdoviridae	*Varicosavirus*	Beet oak leaf virus	BOLV	ssRNA(-)	PV-1034	*Spinacia oleracea*	OQ975887-88
Rhabdoviridae	*Alphanucleorhabdovirus*	Physostegia chlorotic mottle virus	PhCMoV	ssRNA(-)	PV-1182	*Nicotiana occidentalis* “37B”	KX636164
Rhabdoviridae	*Betanucleorhabdovirus*	Sonchus yellow net virus	SYNV	ssRNA(-)	PV-0052	*Nicotiana clevelandii*	MT613317
Aspiviridae	*Ophiovirus*	Lettuce ring necrosis virus	LRNV	ssRNA(-)	PV-0983	*Nicotiana occidentalis* “P1”	ON398506-09
Partitiviridae	*Alphacryptovirus*	Poinsettia latent virus	PnLV	dsRNA	PV-0629	*Euphorbia pulcherrima*	ON398503
Caulimoviridae	*Badnavirus*	Banana streak OL virus	BSOLV	dsDNA-RT	PV-0492	*Musa* sp.	OQ102041
Caulimoviridae	*Caulimovirus*	Cauliflower mosaic virus	CaMV	dsDNA-RT	PV-0229	*Brassica rapa*	OP947586
Geminiviridae	*Begomovirus*	Squash leaf curl virus	SLCV	ssDNA	PV-1299	*Cucurbita pepo*	MW582809-10
Geminiviridae	*Mastrevirus*	Maize streak virus	MSV	ssDNA	PV-1103	*Zea mays*	OQ102042-44
Nanoviridae	*Babuvirus*	Banana bunchy top virus	BBTV	ssDNA	PV-1166	*Musa* sp.	OQ102052-57

(a) DSMZ catalogue code

(b) Host in which the virus isolate was propagated and lyophilized

(c) Several variants are present in the propagated sample and accession numbers for the variants are provided

Table 2. Additional viruses identified by analysis of the HTS data in the samples used to assemble the synthetic mock communities of varying complexity.

Family	Genus	Virus	Acronym	Genome type	Reference sequence accession number[a]
Virgaviridae	*Tobamovirus*	Tobacco mosaic virus	TMV	ssRNA(+)	OQ953825
Tymoviridae	unassigned	Poinsettia mosaic virus	PnMV	ssRNA(+)	OQ953828
Endornaviridae	*Alphaendornavirus*	Hordeum vulgare endornavirus	HvEV	ssRNA(+)	OQ953829
Solemoviridae	*Polerovirus*	Turnip yellows virus	TuYV	ssRNA(+)	JQ862472
Geminiviridae	*Mastrevirus*	Maize streak Réunion virus	MSRV	ssDNA	OQ953826
Totiviridae	unassigned	Maize-associated totivirus	MATV	dsRNA	OQ953827
Totiviridae	unassigned	Maize-associated totivirus 2	MTV-2	dsRNA	MN428829
Mitoviridae	*Duamitovirus*	Chenopodium quinoa mitovirus 1	CqMV1	ssRNA(+)	MT089917
small linear ssRNA satellite		Turnip crinkle satellite RNA F	TCVsatRNA F	ssRNA	X12749
small linear ssRNA satellite		Pea enation mosaic virus satellite RNA	PEMVsatRNA	ssRNA	OQ953831
small linear ssRNA satellite		Strawberry latent ringspot virus satellite RNA	SLRSVsatRNA	ssRNA	OQ953830

(a) Accession number of the closest sequence in GenBank that was used as reference for reads mapping

Table 3. Comparison of the number and average length of *de novo* assembled viral contigs obtained for VANA and dsRNA datasets normalized at different sequencing depths (100K, 300K, 1M, 3M and 10M reads, five resampling repeats at each depth). The standard deviations (SD) and the statistical differences (p-values) are also shown

		VANA average +/- SD	**dsRNA average +/- SD**	**Two sample t-test**
100 K reads	nb viral contigs	33.6 +/- 1.9	101.8 +/- 2.9	*9.2E-11*
	Viral contigs average length	733.4 +/- 23.7	747.4 +/- 17.1	0.32
300K reads	nb viral contigs	70.2 +/- 5.4	129.4 +/- 8.1	*8.0E-07*
	Viral contigs average length	643.4 +/- 27.8	887.8 +/- 38.2	*2.8E-06*
1M reads	nb viral contigs	106.2 +/- 6.3	159.2 +/- 6.6	*1.1E-06*
	Viral contigs average length	694.8 +/- 30.3	1019.6 +/- 40.9	*5.7E-07*
3M reads	nb viral contigs	129.6 +/- 4.8	207.6 +/- 3.8	*2.5E-09*
	Viral contigs average length	798.4 +/- 15.9	1067.6 +/- 11.5	*1.4E-09*
10M reads	nb viral contigs	201.2 +/- 4.1	268 +/- 2.9	*1.8E-09*
	Viral contigs average length	791.2 +/- 11.1	1121.4 +/- 10.6	*3.9E-11*

Table 4. Performance parameters of *de novo* assembly using different minimal contigs length of normalized, 10M reads, VANA and dsRNA datasets for the 60-viruses synthetic community.

	Minimal contig length											
	125 nt		**175 nt**		**250 nt**		**350 nt**		**500 nt**		**1000 nt**	
	VANA	**dsRNA**	**VANA**	**dsRNA**	**VANA**	**dsRNA**	**VANA**	**dsRNA**	**VANA**	**dsRNA**	**VANA**	**dsRNA**
nb contigs	1947	2212	416	784	220	437	144	276	86	182	37	88
average length	235	324	506	607	757	907	985	1243	1355	1662	2191	2696
N25	547	1764	1836	3642	2334	4007	2560	4060	3449	4505	3824	5671
N50	206	352	628	1005	994	1521	1277	1955	1709	2775	2277	3705
N75	156	191	313	352	481	558	618	773	888	1117	1653	1782
Max	6549	13919	6652	13919	6652	13919	6549	13919	6652	13919	6652	13919
nb viral contigs	1852	1672	378	468	204	269	137	181	84	131	37	75
% viral contigs	95%	76%	91%	60%	93%	62%	95%	66%	98%	72%	100%	85%
Viral contigs average length	235	327	525	741	783	1123	1008	1508	1368	1921	2191	2833
Bases in viral contigs	435421	547507	198486	347003	159827	302074	138102	272883	114951	251656	81077	212467
% bases in viral contigs	95.20%	76.40%	94.40%	72.90%	96.00%	76.20%	97.40%	79.50%	98.60%	83.20%	100%	89.50%

Table 5. Detection performance of VANA, dsRNA and RNASeq methods at the level of individual viral genomic molecules (from a total of 115 viral molecules) using the most complex, 60-virus pool, for groups of viruses with different genome types at 10M reads sequencing depth.

	# viral molecules	VANA		dsRNA		RNASeq	
		# detected	% detected	# detected	% detected	# detected	% detected
+ssRNA viruses	86	50	58.1%	77	89.5%	81	96.,3%
-ssRNA viruses	12	1	8.3%	5	41.7%	12	100%
RNA satellites	3	1	33.3%	3	100%	2	66.,6%
dsRNA viruses	2	0	0%	2	100%	0	0,0%
ssDNA viruses	10	4	40.0%	0	0%	4	40.0%
dsDNA viruses	2	1	50.0%	1	50.0%	1	50.0%
Total	**115**	**57**	**49.6%**	**88**	**76.5%**	**100**	**87.,7%**

Supplementary Figures

Supplementary Figure S1: Pooling strategy to generate mock virus communities with different degrees of complexity (5, 10, 20, and 60-viruses communities).

Supplementary Figure S2: Percent coverage of detected viral molecules using the VANA or the dsRNA approaches as a function of minimal contig length.

Supplementary Figure S3: Number of detected RNA viruses based on *de novo* assembled contigs from the VANA or the dsRNA approaches for datasets normalized at a 120K reads sequencing depth and for viral communities with different degrees of complexity.

Supplementary Figure S4: Number of detected RNA viruses for viral communities with different degrees of complexity using *de novo* assembled contigs from the VANA or the dsRNA approaches derived from datasets normalized so as to compensate for community complexity (120K reads for 5 viruses communities, 240K for 10 viruses, 480K for 20 viruses and 1.44M for 60 viruses).

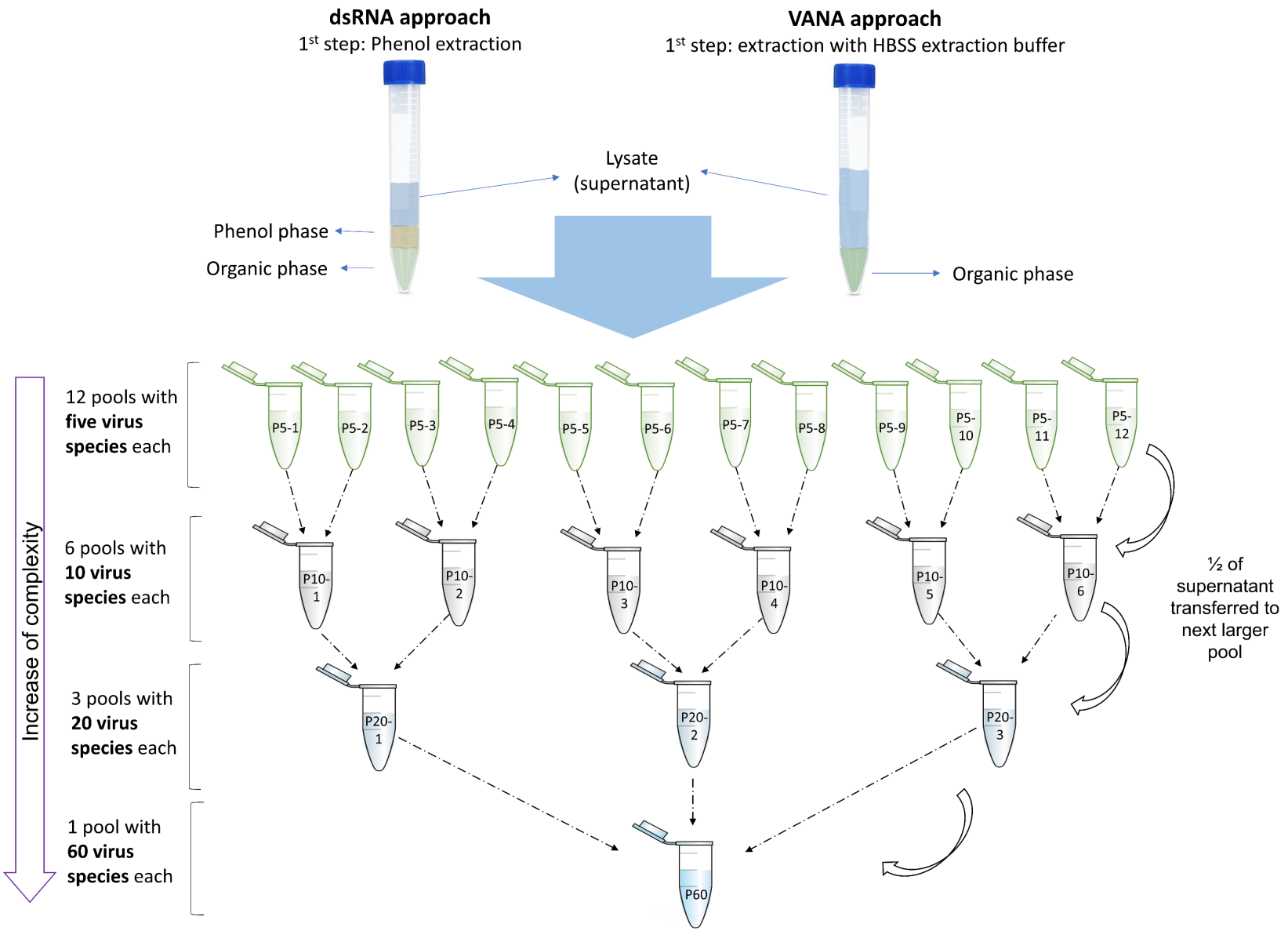
Supplementary Figure S5: Average proportion of the length of viral molecules represented by contigs obtained for the VANA or the dsRNA approaches as a function of sequencing depth. For each sequencing depth 5 independent random resamplings were performed and error bars represent the standard deviations of the coverage obtained.

Supplementary Tables

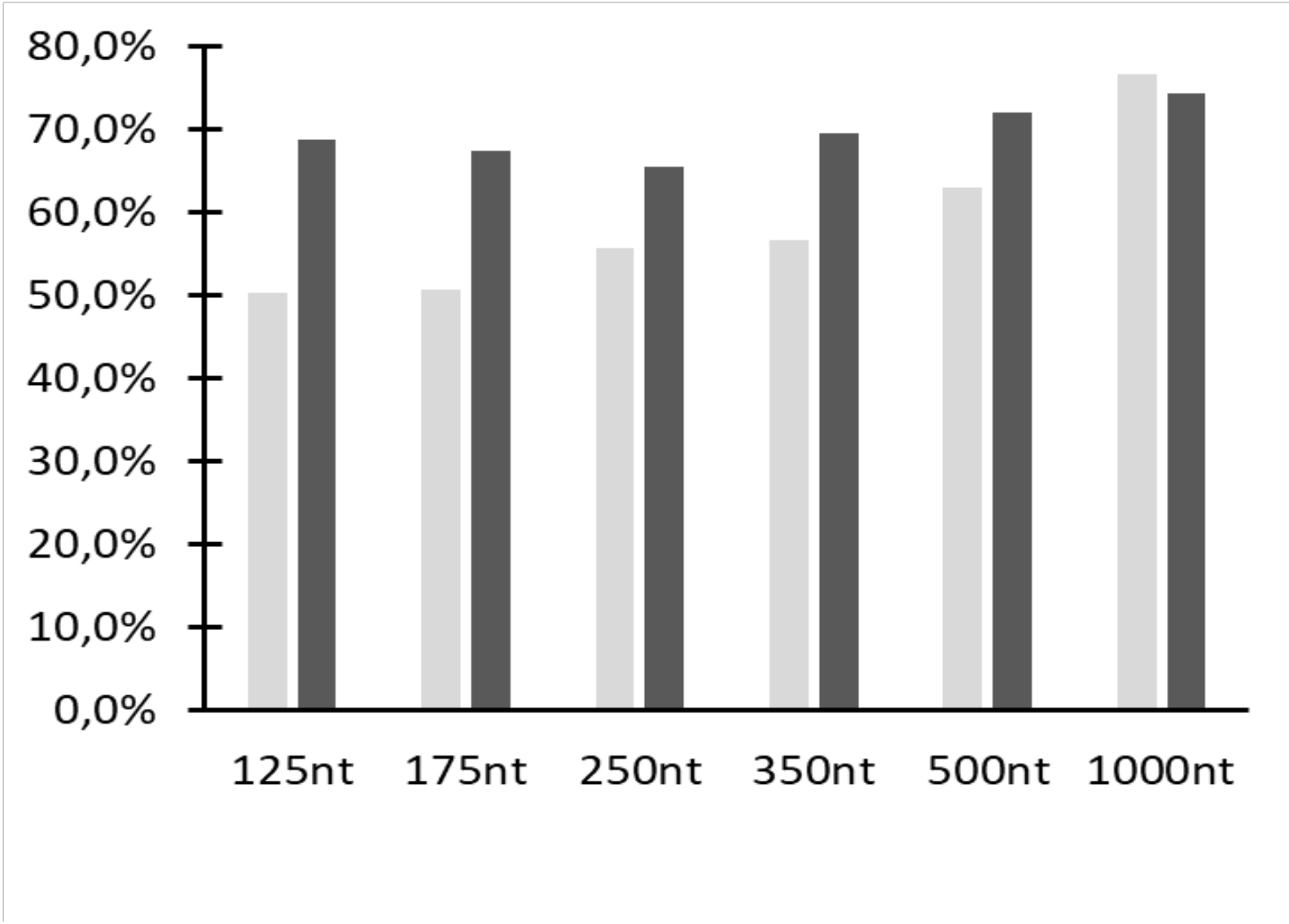
Supplementary Table S1. Pooling strategy to generate the various pools of variable complexity (from 5 to 60 viruses in a pool).

Supplementary Table S2. Comparison of *de novo* assembly parameters for VANA and dsRNA datasets normalized at different sequencing depths (100K, 300K, 1M, 3M and 10M reads, 5 resampling repeats at each depth) and corresponding statistical significance.

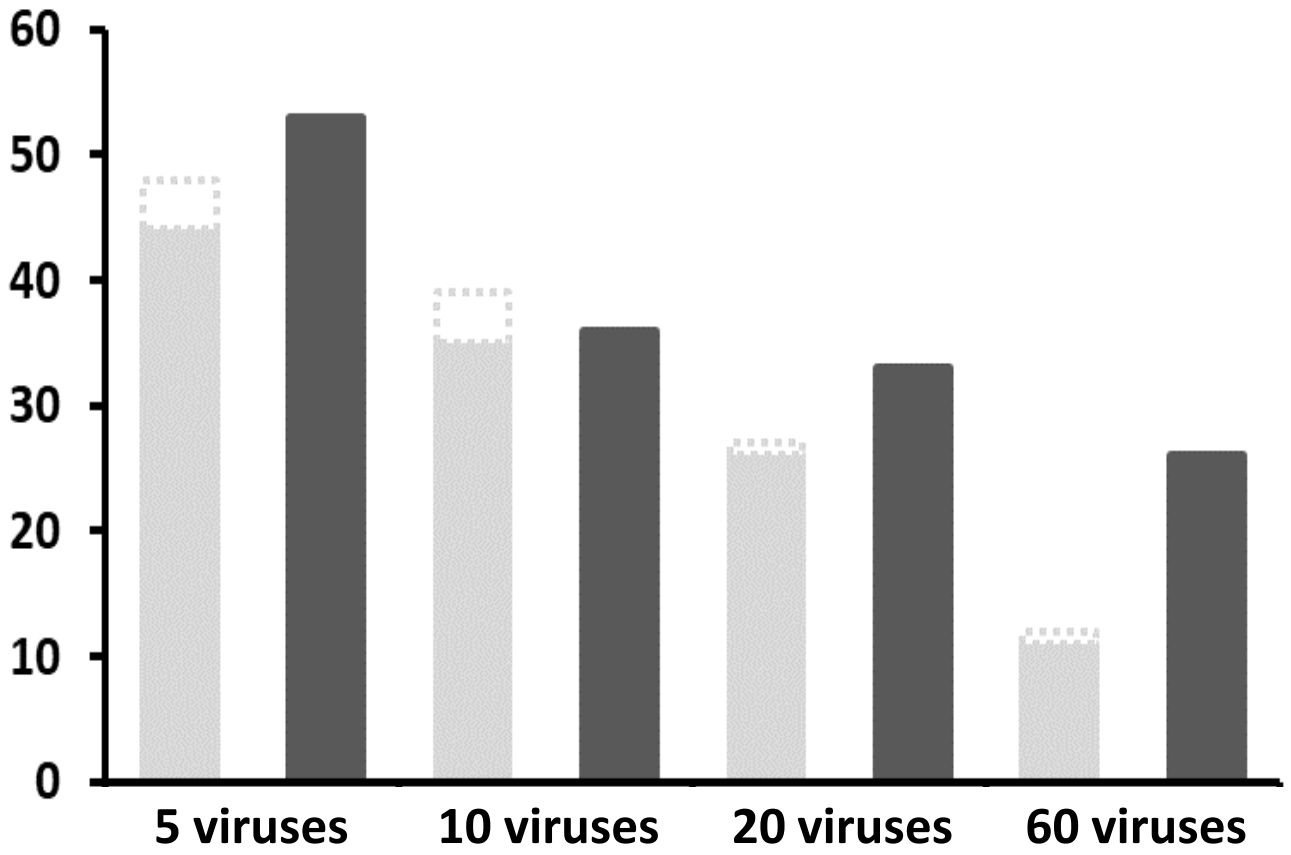
Supplementary Figure S1: Pooling strategy to generate mock virus communities with different degrees of complexity (5, 10, 20, and 60-viruses communities)



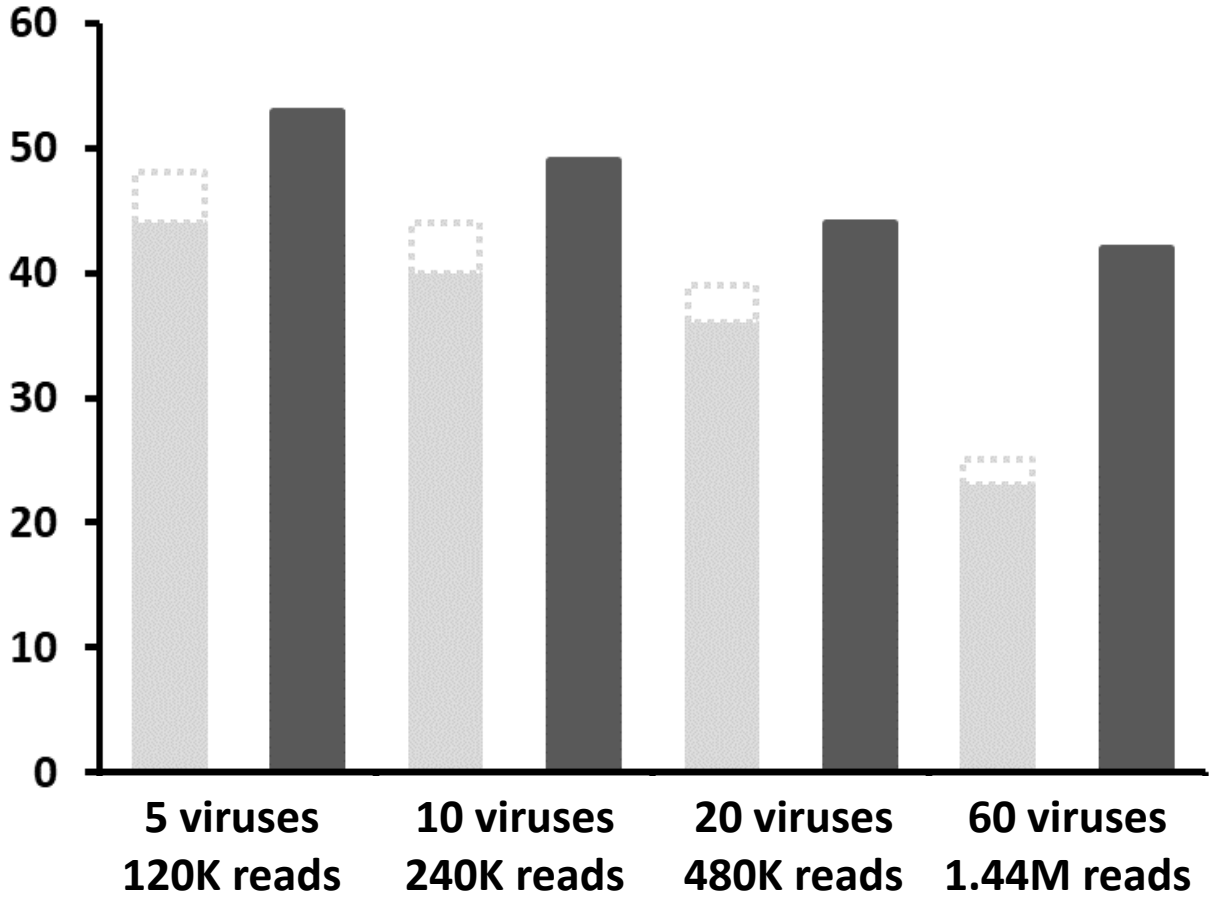
Supplementary Figure S2: Percent coverage of detected viral molecules as a function of minimal contig length for the VANA (light grey) or the dsRNA (dark grey) approaches on the 60 viruses community at a 10 millions reads sequencing depth



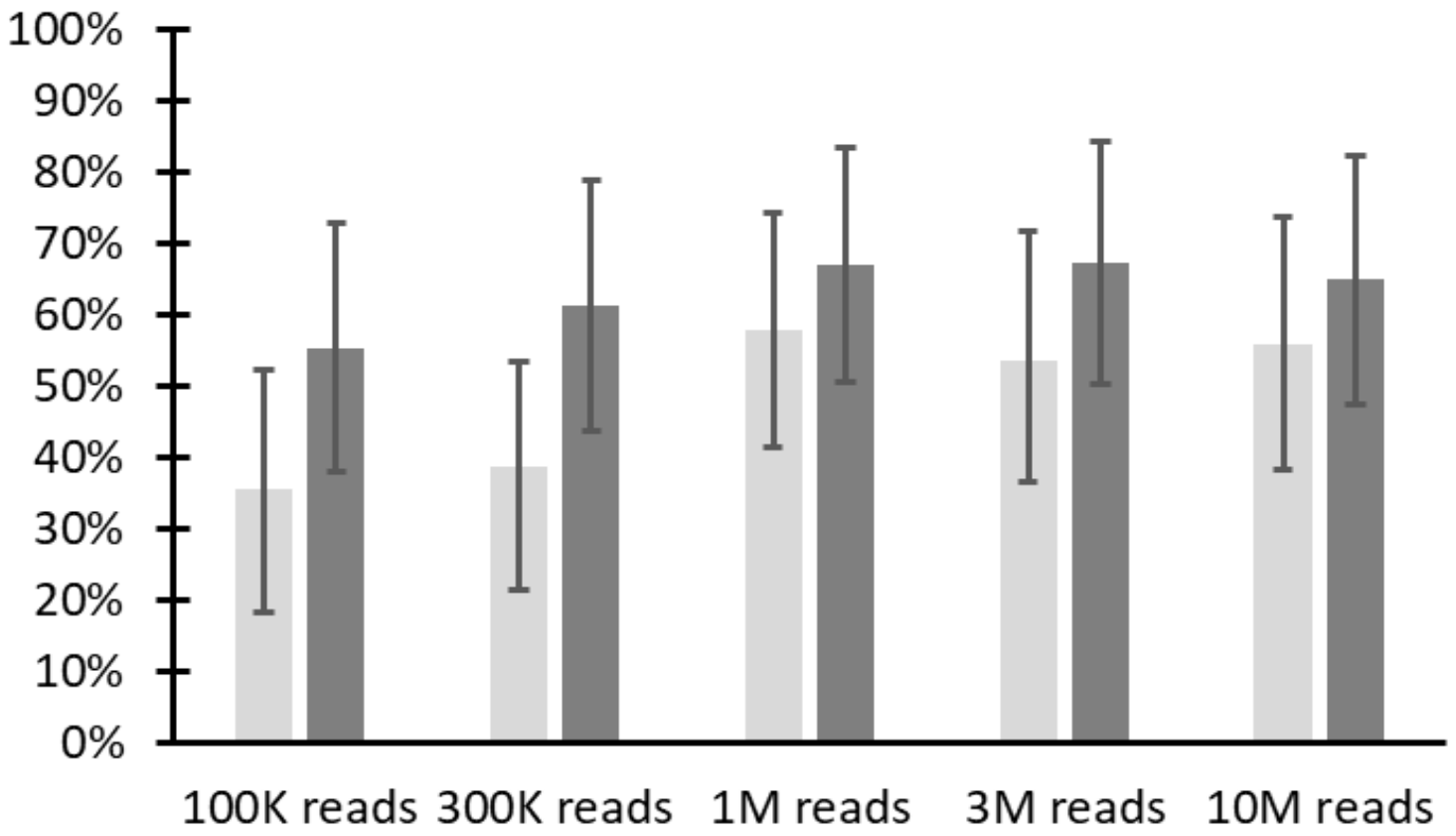
Supplementary Figure S3: Number of detected viruses based on *de novo* assembled contigs from the VANA (light grey) or the dsRNA (dark grey) approaches for datasets normalized at a 120K reads sequencing depth and for viral communities with different degrees of complexity. RNA viruses are indicated by solid bars while DNA viruses are indicated by dashed bars.



Supplementary Figure S4: Number of detected viruses for viral communities with different degrees of complexity using *de novo* assembled contigs from the VANA (light grey) or the dsRNA (dark grey) approaches. Datasets were normalized so as to compensate for community complexity (120K reads for 5 viruses communities, 240K for 10 viruses, 480K for 20 viruses and 1.44M for 60 viruses). RNA viruses are indicated by solid bars while DNA viruses are indicated by dashed bars.



Supplementary Figure S5: Average proportion of the length of detected viral molecules of the 60 viruses community represented by contigs obtained for the VANA (light grey) or the dsRNA (dark grey) approaches at different sequencing depth. For each sequencing depth 5 independent random resamplings were performed and error bars represent the standard deviations of the coverage obtained



Supplementary Table S1: pooling strategy to generate the viral communities pools of variable complexity (from 5 to 60 viruse

60 viruses pool	20 viruses pools	10 viruses pools	5 viruses pools	Family	Genus
P60	P20-1	P10-1	P5-1	*Bromoviridae*	*Alfamovirus*
				Alphaflexiviridae	*Allexivirus*
				Tombusviridae	*Alphacarmovirus*
				Partitiviridae	*Alphacryptovirus*
				Benyviridae	*Benyvirus*
			P5-2	*Tombusviridae*	*Alphanecrovirus*
				Bromoviridae	*Anulavirus*
				Potyviridae	*Bymovirus*
				Caulimoviridae	*Badnavirus*
				Betaflexiviridae	*Capillovirus*
		P10-2	P5-3	*Tombusviridae*	*Betacarmovirus*
				Geminiviridae	*Begomovirus*
				Betaflexiviridae	*Carlavirus*
				Bromoviridae	*Bromovirus*
				Secoviridae	*Cheravirus*
			P5-4	*Nanoviridae*	*Babuvirus*
				Tombusviridae	*Aureusvirus*
				Closteroviridae	*Closterovirus*
				Secoviridae	*Comovirus*
				Solemoviridae/Luteoviridae	*Enamovirus/Umbravirus*
	P20-2	P10-3	P5-5	*Tombusviridae*	*Betanecrovirus*
				Caulimoviridae	*Caulimovirus*
				Closteroviridae	*Crinivirus*
				Bromoviridae	*Cucumovirus*
				Virgaviridae	*Furovirus*
			P5-6	*Rhabdoviridae*	*Cytorhabdovirus*
				Tombusviridae	*Dianthovirus*
				Secoviridae	*Fabavirus*
				Not assigned	*Idaeovirus*
				Virgaviridae	*Hordeivirus*
		P10-4	P5-7	*Tombusviridae*	*Gammacarmovirus*
				Bromoviridae	*Ilarvirus*
				Potyviridae	*Ipomovirus*
				Rhabdoviridae	*Varicosavirus*
				Tombusviridae	*Machlomovirus*
			P5-8	*Geminiviridae*	*Mastrevirus*
				Secoviridae	*Nepovirus*
				Luteoviridae	*Polerovirus*
				Virgaviridae	*Pomovirus*
				Betaflexiviridae	*Trichovirus*
	P20-3	P10-5	P5-9	*Virgaviridae*	*Pecluvirus*
				Tombusviridae	*Pelarspovirus*
				Alphaflexiviridae	*Potexvirus*
				Potyviridae	*Potyvirus*
				Rhabdoviridae	*Alphanucleorhabdovirus*
			P5-10	*Rhabdoviridae*	*Betanucleorhabdovirus*
				Potyviridae	*Rymovirus*
				Secoviridae	*Sequivirus*
				Solemoviridae	*Sobemovirus*
				Betaflexiviridae	*Tepovirus*
		P10-6	P5-11	*Aspiviridae*	*Ophiovirus*
				Virgaviridae	*Tobravirus*
				Tombusviridae	*Tombusvirus*
				Potyviridae	*Tritimovirus*
				Tymoviridae	*Tymovirus*
			P5-12	*Virgaviridae*	*Tobamovirus*
				Secoviridae	*unassigned*
				Tombusviridae	*Umbravirus*
				Potyviridae	unassigned
				Tospoviridae	*Orthotospovirus*

s in a pool)

Virus species	Genome type
Alfalfa mosaic virus	ssRNA(+)
Shallot virus X	ssRNA(+)
Calibrachoa mottle Virus	ssRNA(+)
Poinsettia latent virus	dsRNA
Beet necrotic yellow vein virus	ssRNA(+)
Tobacco necrosis virus A	ssRNA(+)
Pelargonium zonate spot virus	ssRNA(+)
Barley yellow mosaic virus	ssRNA(+)
Banana streak OL virus	dsDNA-RT
Apple stem grooving virus	ssRNA(+)
Turnip crinkle virus	ssRNA(+)
Squash leaf curl virus	ssDNA
Poplar mosaic virus	ssRNA(+)
Brome mosaic virus	ssRNA(+)
Arracacha virus B	ssRNA(+)
Banana bunchy top virus	ssDNA
Johnsongrass chlorotic stripe mosaic virus	ssRNA(+)
Beet yellows virus	ssRNA(+)
Squash mosaic virus	ssRNA(+)
Pea enation mosaic virus 1 and 2	ssRNA(+)
Beet black scorch virus	ssRNA(+)
Cauliflower mosaic virus	dsDNA-RT
Tomato chlorosis virus	ssRNA(+)
Peanut stunt virus	ssRNA(+)
Soil-borne wheat mosaic virus	ssRNA(+)
Lettuce necrotic yellows virus	ssRNA(-)
Carnation ringspot virus	ssRNA(+)
Broad been wilt virus 1	ssRNA(+)
Raspberry bushy dwarf virus	ssRNA(+)
Barley stripe mosaic virus	ssRNA(+)
Melon necrotic spot virus	ssRNA(+)
Parietaria mottle virus	ssRNA(+)
Cucumber vein yellowing virus	ssRNA(+)
Beet oak leaf virus	ssRNA(-)
Maize chlorotic mottle virus	ssRNA(+)
Maize streak virus	ssDNA
Tomato black ring virus	ssRNA(+)
Cucurbit aphid-borne yellows virus	ssRNA(+)
Potato mop-top virus	ssRNA(+)
Apple chlorotic leaf spot virus	ssRNA(+)
Peanut clump virus	ssRNA(+)
Pelargonium line pattern virus	ssRNA(+)
Lettuce virus X	ssRNA(+)
Bidens mottle virus	ssRNA(+)
Physostegia chlorotic mottle virus	ssRNA(-)
Sonchus yellow net virus	ssRNA(-)
Agropyron mosaic virus	ssRNA(+)
Carrot necrotic dieback virus	ssRNA(+)
Rice yellow mottle virus	ssRNA(+)
Potato virus T	ssRNA(+)
Lettuce ring necrosis virus	ssRNA(-)
Pea early-browning virus	ssRNA(+)
Tomato bushy stunt virus	ssRNA(+)
Brome streak mosaic virus	ssRNA(+)
Turnip yellow mosaic virus	ssRNA(+)
Paprika mild mottle virus	ssRNA(+)
Strawberry latent ringspot virus	ssRNA(+)
Carrot mottle virus	ssRNA(+)
Spartina mottle virus	ssRNA(+)
Impatiens necrotic spot virus	ssRNA(+/-)-

Supplementary Table S2. Comparison of *de novo* assembly parameters for VANA and dsRNA datasets normalized at different sequencing depths (100K, 300K, 1M, 3M and 10M reads, 5 resamplings at each depth) and corresponding statistical significance.

		VANA average +/- SD	dsRNA average +/- SD	Two sample t-test
100K reads	nb contigs	33.6 +/- 1.9	103.2 +/- 2.4	*2.6E-11*
	average length	733.4 +/- 23.7	741.8 +/- 14.9	0.52
	N50	839.2 +/- 104.8	937.2 +/- 34.1	0.10
	Max length	5886.2 +/- 442.4	5277.2 +/- 577.9	0.10
	nb viral contigs	33.6 +/- 1.9	101.8 +/- 2.9	*9.2E-11*
	% viral contigs	100% +/- 0%	99% +/- 1%	*2.5E-02*
	Viral contigs average length	733.4 +/- 23.7	747.4 +/- 17.1	0.32
300K reads	nb contigs	70.6 +/- 6.1	140.4 +/- 8.9	*5.1E-07*
	average length	642.4 +/- 30.0	849.4 +/- 35.3	*8.5E-06*
	N50	681.2 +/- 93.1	1102.4 +/- 104.5	*1.5E-04*
	Max length	6260 +/-213.1	7581.6 +/- 3875.5	0.49
	nb viral contigs	70.2 +/- 5.4	129.4 +/- 8.1	*8.0E-07*
	% viral contigs	100% +/- 1%	92% +/- 2%	*7.7E-05*
	Viral contigs average length	643.4 +/- 27.8	887.8 +/- 38.2	*2.8E-06*
1M reads	nb contigs	108.2 +/- 6.1	198.4 +/- 7.8	*3.5E-08*
	average length	687.6 +/- 28.1	915.4 +/- 35.4	*3.5E-06*
	N50	767 +/- 74.9	1556.4 +/- 178.3	*2.6E-04*
	Max length	6491 +/- 79.4	10382.6 +/- 1825.7	*8.9E-03*
	nb viral contigs	106.2 +/- 6.3	159.2 +/- 6.6	*1.1E-06*
	% viral contigs	98% +/- 1%	80% +/- 2%	*8.3E-08*
	Viral contigs average length	694.8 +/- 30.3	1019.6 +/- 40.9	*5.7E-07*
3M reads	nb contigs	134.2 +/- 3.4	284 +/- 8.2	*2.5E-07*
	average length	783.4 +/- 10.1	931 +/- 16.6	*1.5E-07*
	N50	1016.6 +/- 73.5	1599.4 +/- 214.1	*4.3E-04*
	Max length	6540 +/- 3.4	11573.4 +/- 2359.2	*8.8E-03*
	nb viral contigs	129.6 +/- 4.8	207.6 +/- 3.8	*2.5E-09*
	% viral contigs	97% +/- 1%	73% +/- 1%	*1.0E-09*
	Viral contigs average length	798.4 +/- 15.9	1067.6 +/- 11.5	*1.4E-09*
10M reads	nb contigs	217 +/- 4.4	433.8 +/- 4.4	*8.2E-13*
	average length	764.6 +/- 10.5	907.2 +/- 7.7	*8.2E-09*
	N50	1025.4 +/- 31.2	1529.6 +/- 21.7	*1.8E-09*
	Max length	6534.4 +/- 210.2	13930.4 +/- 24.9	*1.6E-07*
	nb viral contigs	201.2 +/- 4.1	268 +/- 2.9	*1.8E-09*
	% viral contigs	93% +/- 0%	62% +/- 1%	*2.2E-13*
	Viral contigs average length	791.2 +/- 11.1	1121.4 +/- 10.6	*3.9E-11*