BMC Genomics

RESEARCH

Open Access

From reads to results: comparing Oxford Nanopore to Illumina sequencing for citrus virus surveillance

Madelein Dippenaar[1], Hans Jacob Maree[1,2] and Rachelle Bester[1,2*]

Abstract

Background ONT sequencing has been previously evaluated for its ability to detect plant viruses and viroids. Its advantages, such as longer read lengths and real-time analysis, compete with extensively validated Illumina platforms for possible incorporation into routine pathogen detection. The continuous development and improvement of ONT sequencing along with the discontinuation of older equipment and reagents, necessitate a renewed comparison of these sequencing platforms, specifically in citrus where only limited research is available.

Results This study compared the ability of Oxford Nanopore Technologies (ONT) sequencing, using the MinION flow cell, with Illumina sequencing, on the NovaSeqX platform, to accurately detect three viruses and three viroids in citrus. Both technologies were able to identify all pathogens using both reference-dependent and independent methods. While Illumina sequencing re-established the high sensitivity, coverage and accuracy seen previously, ONT compensated for fewer pathogen reads and lower depth with longer reads that enabled reasonable genome coverage and sequencing identities comparable to that of Illumina. Moreover, the pooling of data from different ONT barcode datasets from a single sample, improved comparability to Illumina results as small variations in library preparation, sample loading and flow cells can lead to a significant decrease in sequencing data. Reference gene expression profiles were also investigated to evaluate internal controls and check outlier samples. The ONT platform also had a shorter turnaround compared to Illumina sequencing.

Conclusion The use of ONT sequencing may offer advantages for small-scale routine pathogen detection. It has the potential to accurately detect pathogens and discover novel viral agents. This comparison between Illumina and ONT platforms highlights the strengths associated with each approach and offers new insights into the possible application of high-throughput sequencing (HTS) within plant health surveillance and biosecurity programs.

Keywords High-throughput sequencing, RNA viruses, Viroids, Pathogen identification, MinION, NovaSeqX.

*Correspondence:
Rachelle Bester
rachelle@sun.ac.za
[1]Department of Genetics, Stellenbosch University, Stellenbosch, South Africa
[2]Citrus Research International, Stellenbosch, South Africa

Background

High-throughput sequencing (HTS) technologies have altered the landscape of pathogen identification and detection by offering various advantages over conventional methods. It has the ability to detect novel pathogens, analyse and elucidate mixed infections simultaneously, offer new insights into pathogen biology and ecology as well as the potential for accurate and rapid diagnostics. This makes it an attractive technology to implement in viral disease management [1–6]. Various agriculturally important crop industries are globally threatened by an array of viral pathogens. Due to the nature of viral pathogen transmission; surveillance, certification and quarantine programs are crucial to prevent the introduction and dissemination of potentially harmful pathogens. The cornerstone of these programs is reliable, accurate and, to a certain extent, rapid pathogen detection [1, 7, 8]. These programs could, therefore, greatly benefit from the inclusion of HTS and various countries are currently in different stages of incorporating this technology into their plant disease management protocols [2, 4, 9, 10]. Given the relative recent development of this technology, the issues regarding standardisation and optimisation for routine implementation thereof persist. The difficulty in adopting HTS for these programs is largely because the technology is still evolving. New developments and improvements occur constantly while older reagents, instruments, software and analysis pipelines are discontinued. As a result, established protocols must be continually updated and adapted [2, 8, 11].

Consequently, multiple platforms based on different sequencing chemistries have been developed and validated for plant viral pathogen diagnostic use [1, 3, 5, 6, 9, 12]. Notably, the most widely used platform is Illumina, which provides high-throughput, short-reads with very low error rates [3, 10]. Recently, third-generation technologies like nanopore sequencing from Oxford Nanopore Technologies (ONT) have come to rival the performance of Illumina by offering alternative advantages while maintaining an adequate sequencing quality. ONT generates long-read sequences that are preferred for certain applications and it has the added benefit of possible real-time analysis [13, 14]. Furthermore, ONT has scalable options for sequencing including the portable MinION device that can be ideal for use in small laboratories and possible in-field applications [8, 15]. This variation in sequencing options also implies that the amount of data generated varies. Compared to Illumina, ONT produces a relatively large error rate [14], although according to the newest release from ONT, the kit V14 chemistries used in conjunction with R10.4.1 flow cells enable over 99% (Q20) accuracy across all read lengths [16].

The comparison between Illumina and ONT for application in plant virology is currently fairly limited, even more so for viroid detection, especially given the fact that most of the consumables used in previous studies have been replaced and/or discontinued [5, 6, 8, 11, 15, 16]. A study conducted by Pecman et al. in 2022 [8] was the first to use ONT for viroid sequencing. The ONT MinION sequencer was compared with the Illumina MiSeq benchtop sequencer using tobacco plants with established mixed infections of viruses and/or viroids. This study not only compared sequencing platforms but also included a comparison of different ONT sequencing approaches. They found that cDNA-PCR sequencing of rRNA-depleted total RNA was most comparable with Illumina results. Main findings highlighted that while ONT can produce a similar percentage of virus reads in the dataset, Illumina still outperformed it regarding viroid detection. The authors also emphasised that with an adequate amount of data generated, high quality viral genomes with high sequence identity could be reconstructed. These results regarding the efficacy of cDNA-based ONT sequencing approaches and good genome coverage for some viral pathogens have also been corroborated in additional studies on apples, citrus and grapevine [5, 6, 11]. Contrary to this, in a study conducted on wheat in 2023 [15], sequencing on the MinION device with a Flongle adapter showed lower viral genome coverage and sequence identity compared to data generated by the Illumina Hiseq 4000. This was attributed to low data throughput and high error rates [15] which is likely due to the use of the Flongle flow cells which produce less sequencing data than MinION flow cells. Concerns were also expressed regarding ONT's ability to detect viruses at low titer [15]. Recently, in a study on apples, the MinION and Illumina NextSeq2000 were considered, and this particular study investigated seven cDNA synthesis treatments preceded by a template-switching cDNA amplification protocol [11]. While relative ONT virus abundances were higher when certain cDNA synthesis treatments were used, the Illumina platform still generated more viroid reads [11]. Interestingly, the treatments that produced the most favourable results for viruses and viroids, were not the same indicating how differences in pathogen genomes might also influence detection with ONT.

As previously mentioned, ONT has also been used for citrus virus identification [5, 17, 18], most notably for citrus tristeza virus (CTV) [5, 18]. Citrus is a globally important fruit crop [7] with South Africa designated as the second largest exporter of fresh citrus [1] making it an important crop to investigate HTS for pathogen detection. Including a range of pathogens is imperative for a panoptic comparison of sequencing platforms [8] and citrus provides a natural host for various viruses and viroids

[1, 7, 19]. In the present study, three plants were infected with three viruses; CTV, citrus virus A (CiVA) and citrus tatter leaf virus (CTLV); and three viroids; citrus exocortis viroid (CEVd), citrus dwarfing viroid (CDVd) and hop stunt viroid (HSVd). These plants were subjected to Illumina (NovaSeqX) and ONT (MinION) sequencing. Data generated from these platforms were critically compared based on their ability to accurately detect all pathogens using reference-dependent and independent bioinformatics approaches. All of these pathogens are currently incorporated into the citrus improvement scheme of South Africa and results obtained in this study could directly contribute towards the validation of the use of HTS in not only citrus but various plant quarantine and certification programs. Comparing Illumina and ONT strategies and carefully evaluating all factors surrounding these platforms, may help diagnosticians find the middle ground between performance and practicality in plant virus diagnostics.

Methods

Plant material and total nucleic acid extraction

Three *Citrus sinensis* (L.) Osbeck cv. 'Madam Vinous' sweet orange plants with established infections were selected to perform the comparison between ONT and Illumina sequencing platforms (MV4, MV5 and MV9). These plants were previously graft inoculated with a source plant with a pathogen profile well-characterised by both HTS and RT-PCR [1]. This virus profile included CTV with a 19.3 kb positive-sense single-stranded RNA genome (+ ssRNA) [20], CiVA with a negative sense single-stranded (-ssRNA) bipartite genome of 6.5 kb (RNA1) and 2.3 kb (RNA2) [21] and CTLV which is 6.5 kb and also +ssRNA [7]. The viroid profile included CDVd (297 nt), CEVd (372 nt) and HSVd (302 nt).

Total nucleic acid was isolated from the leaf midribs and petioles using a protocol adapted from Arruabarrena et al. (2016) and Oñate-Sánchez and Vicente-Carbajosa (2008) [22, 23]. Most modifications made included the upscaling of the extraction protocol to accommodate a larger amount of input plant material (0.457 ± 0.138 g) but the content and concentration of all buffers and solutions remained the same. Briefly, plant material was homogenised in 5 ml cell lysis solution (pH 4) in a Bioreba extraction bag and incubated at room temperature for 5 min. Approximately 2 ml plant liquid was transferred to clean microcentrifuge tubes and centrifuged (17 136 relative centrifugal force (rcf)) at room temperature for 1 min to pellet debris. 1.2 ml of the supernatant was transferred to a new microcentrifuge tube and 400 µl of protein-DNA-precipitation solution was added and microcentrifuge tubes were inverted to mix contents. These tubes were then incubated on ice for 10 min, and centrifuged (17 136 rcf) for an hour at 4 °C. The supernatant was then transferred to new microcentrifuge tubes (1 ml) and to precipitate total nucleic acid, 1x isopropanol was added to each sample and centrifuged (17 136 rcf) for 15 min at 4 °C. After precipitation, the supernatant was poured off and the pellet was washed with 600 µl cold 70% ethanol and centrifuged (17 136 rcf) for 1 min at 4 °C. Ethanol was removed and the nucleic acid pellet was resuspended in 100 µl Milli-Q® water (Merck). Total nucleic acid was extracted from each plant once and the same extract was used for both Illumina and ONT sequencing.

HTS sequencing

A ribo-depleted RNA library for Illumina sequencing was constructed for each sample with the Illumina TruSeq Stranded Total RNA Sample Preparation kit with Plant Ribo-Zero at Macrogen (South Korea). Paired-end HTS (2 × 150 bp) was performed on an Illumina NovaSeqX instrument (Macrogen, South Korea).

ONT sequencing preparation was divided into three main steps: ribosomal RNA depletion, 3' polyadenylation of RNA (poly-A tailing) and then finally library preparation. Ribo-depletion of samples was performed with the RiboMinus™ Plant Kit for RNA-seq (ThermoFisher Scientific). The polyadenylation of samples was carried out as described in the protocol provided by Oxford Nanopore Technologies [24] which included the use of E. coli poly(A) polymerase (New England Biolabs). The cDNA-PCR Barcoding Kit 24 V14 kit (SQK-PCB114.24, Oxford Nanopore Technologies) was used for library construction (according to the manufacturer's instructions). Each sample was barcoded with four different barcodes, (MV4 with Barcode 1–4, MV5 with Barcode 5–8 and MV9 with Barcode 9–12). The barcodes were pooled to ensure a final library for each sample. Each pooled library was then sequenced on a separate flow cell (R10.4.1) for 72 h using the MinION device (Oxford Nanopore Technologies) and MinKNOW 24.11.10 software. Base-calling and demultiplexing were carried out with the MinKNOW software using the super-accurate base-calling (SUP) 4.3.0 model. For each sample (plant) individual datasets associated with each barcode as well as the combined barcodes per sample, which consisted of the sample data without demultiplexing, were analysed. Each sample was barcoded using four barcodes to simulate sample multiplexing to evaluate the impact on data generation.

Sequencing quality assessment and trimming

The Illumina sequencing quality was assessed using FastQC 0.12.1 [25] and SeqKit 2.7.0 [26]. Adapter sequence removal and quality trimming were performed using Trimmomatic 0.39 [27] (SLIDINGWINDOW of 3 nts with Q20, MINLEN of 20 nts) [28]. A quality threshold of Q10 was applied to ONT data as part of the

base-calling step using the SUP model of the MinKNOW sequencing setup. Similar to the Illumina data, the ONT data was subsequently first visually assessed with NanoQC 0.9.4 and then basic sequencing statistics were obtained with NanoStat 1.6.0 [29]. Quality trimming of the ONT data was carried out with NanoFilt 2.8.0 (minimum length of 20 and minimum average read quality score of 10) [29]. The poly-A tails of reads, generated during library preparation, were removed using Cutadapt 1.9.1 [30].

De novo assembly and taxonomic sequence classification

Trimmed Illumina data was subjected to *de novo* assembly using SPAdes 3.14 and default parameters [31]. The assembled contigs were identified using BLAST+ standalone against a local copy of the NCBI GenBank nucleotide database (downloaded January 2025) using the Blastn 2.15.0 algorithm [1, 28]. Given the long reads generated with ONT sequencing and the relatively small genome size of viroids, trimmed and unassembled ONT reads were identified using Blastn following the same identification procedure used for the Illumina data. For further comparison, Kraken2 2.1.3 taxonomic sequence classification [32] was performed on Illumina and ONT datasets using the unassembled trimmed data for both platforms. Applicable graphs were constructed using R available from the Comprehensive R Archive Network (CRAN) [33] and the package ggplot2 [34].

Electronic diagnostic nucleic acid analysis

Trimmed data from Illumina and ONT platforms were subjected to Electronic Diagnostic Nucleic Acid Analysis (EDNA) on the MiFi® bioinformatics platform [35]. Suitable electronic probes (e-probes) were selected from the e-probe database available on the platform. The e-probes (Additional file 1) were selected based on the pathogen profile of the source plant but also included additional probes for all possible citrus viroids.

Read mapping

The trimmed reads of Illumina and ONT datasets were mapped to reference genomes obtained from Genbank corresponding to the pathogen profile of the source plant (Additional file 2). In order to have an adequate representation of CTV diversity, eight CTV genotypes were included in the read mapping analysis [28]. The trimmed datasets were also mapped to a set of 12 genes, retrieved from Genbank (Additional file 2), that were previously evaluated as *C. sinensis* reference genes [36]. Pathogen and reference gene mapping were performed separately but followed the same general pipeline. For Illumina data, the Burrows-Wheeler Alignment Tool (BWA) 0.7.18 was utilised [37] and the ONT data was mapped using minimap2 2.26 [38, 39]. All mapped reads were filtered for 80% similarity over 50% of the read and the coverage was determined using samtools 1.20 [28, 40]. Additionally, for the pathogen genome mapping, the per base (position) read depth was calculated with samtools and the consensus sequences were extracted (samtools and bcftools 1.19) [41], aligned with MAFFT 7.49 [42] and the sequence identities between Illumina and ONT consensus sequences determined. The minimum depth cutoff was set at one, positions with zero coverage where masked (by using "-") and IUPAC nucleotide ambiguity codes were used where appropriate.

In order to make comparisons between samples and platforms, the read count of each reference gene and pathogen was normalised for biological and technical variation by calculating the transcripts per million (TPM) [1]. This was done separately for each mapping instance. To determine the TPM of a sample, the read count per kilobase of sequence (RPK) was calculated for every reference sequence. The sum of all the RPKs was then divided by a million to yield the denominator for the TPM calculation. The TPM of each reference sequence was then determined by dividing the sequence's RPK with the specific denominator. Where appropriate, comparisons between ONT and Illumina data were substantiated with statistical analyses using R [33]. Significance testing was performed separately for each reference (genes and pathogens) with the Kruskal–Wallis rank sum non-parametric test [43, 44] followed by multiple comparison testing (n = 3) with Dunn's test using Bonferroni correction [45, 46]. All graphs were constructed using the ggplot2 package [34].

Sequencing cost calculation

The sequencing cost for a single sample was estimated for Illumina and ONT sequencing platforms. Quotes and prices that are pertinent to the specific research environment were used in a context-specific platform cost evaluation. Illumina sequencing was carried out by Macrogen (South Korea), and the cost of sequencing was calculated with the quoted amount provided by the service provider. For ONT sequencing, the cost of the amount of each reagent necessary for a sequencing reaction with only one barcode on a single flow cell was determined. This comparison between platforms comprised of the relevant reagents and consumables and not equipment-associated expenses.

Results

Sequencing data quality

Illumina sequencing generated on average 24 million paired-end reads with an average of 3.64 gigabases (Gb) per sample (Table 1). This platform yielded exceptional sequencing quality with an average data retention of 96% per sample after quality trimming. ONT sequencing

Table 1 Sequencing quality metrics for Illumina and Oxford Nanopore Technologies (ONT) datasets

	MV4				**MV5**				**MV9**			
	Raw Reads		**Trimmed Reads**		**Raw Reads**		**Trimmed Reads**		**Raw Reads**		**Trimmed Reads**	
Illumina												
	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2
Mean read length	151	151	148.1	147.8	151	151	148.1	147.6	151	151	146.8	145.7
Mean read quality	34.49	34.24	34.92	34.92	34.79	34.43	35.24	35.12	34.16	32.58	34.87	34.00
Total reads	2.0E+07	2.0E+07	2.0E+07	2.0E+07	2.4E+07	2.4E+07	2.3E+07	2.3E+07	2.8E+07	2.8E+07	2.7E+07	2.7E+07
Total bases	3.0E+09	3.0E+09	2.9E+09	2.9E+09	3.6E+09	3.6E+09	3.5E+09	3.5E+09	4.3E+09	4.3E+09	4.0E+09	4.0E+09
Oxford Nanopore Technologies (ONT)												
	Combined				Combined				Combined			
Mean read length	693.8		668.2		634		613.2		537.6		528.5	
Mean read quality	13		13.1		12.8		12.9		12.4		12.5	
Total reads	1.3E+07		1.2E+07		6.0E+06		5.5E+06		1.8E+07		1.6E+07	
Total bases	9.1E+09		8.2E+09		3.8E+09		3.4E+09		9.5E+09		8.6E+09	
	Barcode 1				Barcode 5				Barcode 9			
Mean read length	682.8		659.4		601		579.5		542.9		532.8	
Mean read quality	12.9		13		12.7		12.8		12.4		12.6	
Total reads	2.6E+06		2.5E+06		1.6E+06		1.5E+06		4.9E+06		4.6E+06	
Total bases	1.8E+09		1.6E+09		9.8E+08		8.7E+08		2.7E+09		2.4E+09	
	Barcode 2				Barcode 6				Barcode 10			
Mean read length	696.2		667.3		708.3		683.9		551.5		544	
Mean read quality	13.2		13.3		13.2		13.3		12.5		12.6	
Total reads	3.7E+06		3.5E+06		8.3E+05		7.8E+05		3.6E+06		3.3E+06	
Total bases	2.6E+09		2.3E+09		5.9E+08		5.3E+08		2.0E+09		1.8E+09	
	Barcode 3				Barcode 7				Barcode 11			
Mean read length	698.6		669.5		630.6		609.6		562.9		552.9	
Mean read quality	13.1		13.2		12.8		13		12.7		12.8	
Total reads	3.5E+06		3.3E+06		1.6E+06		1.4E+06		3.1E+06		2.9E+06	
Total bases	2.4E+09		2.2E+09		9.9E+08		8.8E+08		1.8E+09		1.6E+09	
	Barcode 4				Barcode 8				Barcode 12			
Mean read length	694.8		674.8		632.4		613.7		511.8		502.7	
Mean read quality	12.9		13.1		12.7		12.8		12.2		12.3	
Total reads	3.3E+06		3.1E+06		1.9E+06		1.8E+06		6.1E+06		5.5E+06	
Total bases	2.3E+09		2.1E+09		1.2E+09		1.1E+09		3.1E+09		2.8E+09	

quality was assessed individually for each barcode dataset (MV4 Barcode 1–4, MV5 Barcode 5–8 and MV9 Barcode 9–12) as well as the combined barcode dataset for each sample (Table 1). Each sample was sequenced on a separate flow cell which varied with regards to pore activity at the start of sequencing. The sequencing run of MV4 had 1279 available pores at the start of sequencing and yielded 9.10 Gb of raw combined data, MV5's run had considerably fewer active pores (815) and generated 3.78 Gb and MV9's flow cell performed similarly to the first sequencing run with 1053 pores that produced 9.54 Gb. Overall, a single barcode generated an average of 3 million long reads per sample which corresponded to an average of 1.87 Gb. After quality trimming and removal of poly-A tails, 90% of the ONT data was retained.

De novo assembly and blast

De novo assembled contigs from the Illumina dataset and trimmed ONT datasets were subjected to Blastn analysis to determine the virus and viroid component of each (Table 2). Clear differences in trends were observed between ONT and Illumina. Illumina had a higher average fraction of both viruses (6.9%) and viroids (0.016%) compared to ONT (0.15% and 0.002%). MV9 had a larger proportion of virus and viroid contigs/reads for both Illumina and ONT compared to the other plants, although the composition profile still followed the general trends observed for each sequencing platform. Overall, the various datasets from each plant were very consistent with one another when considering ONT but the MV5 Barcode 7 dataset had a larger proportion of unidentified reads than the other ONT datasets and the MV5 Barcode 6 dataset was the only dataset in which no Blastn hits were attributed to HSVd.

Kraken2 sequence classification

Similar to Blastn, Kraken2 also provided an indication of the proportion of sequences attributed to the host genome and various pathogens through taxonomic sequence classification (Table 3). On average 78% of trimmed Illumina reads and 74% of ONT trimmed reads were identified as host genome. Although the Illumina data had more unclassified contigs (5.6%), it also had a larger proportion on average of virus (8%), and viroid (0.02%) sequences compared to the ONT data that had 0.09% virus and 0.002% viroid sequences. Notably, Illumina also had a considerably larger percentage of sequences attributed to bacteria (0.24%) and fungi (0.01%). Across different plants, the Illumina datasets exhibited high concordance. In contrast, the nanopore datasets showed greater variability, with MV9 differing from MV4 and MV5 due to a lower proportion of reads assigned to the host genome. Additionally, all samples had sequences classified to the species level corresponding to the complete viral and viroid profile expected, except the MV5 Barcode 6 dataset where no sequences were classified as HSVd.

Electronic diagnostic analysis with MiFi®

All e-probes that correspond to the expected pathogen profile indicated a positive diagnostic call on the MiFi® bioinformatics platform (Additional file 3). The probe set designed to detect multiple citrus viroids also indicated a positive call owing to the ability of this probe set to detect CDVd. The additional viroid e-probes (CVd V, CVd VI and CVd VII) were negative for all samples as expected. Similar to the Kraken2 results, no reads associated with HSVd was present in dataset MV5 Barcode 6.

Table 2 Blastn results from *de novo* assembled illumina data En trimmed Oxford nanopore technologies (ONT) data across barcode datasets

Classification*	Percentage of contigs / sequences classified (%)					
MV4	Barcode 1	Barcode 2	Barcode 3	Barcode 4	Combined	Illumina
Virus	0.0922	0.0947	0.0900	0.0882	0.0913	5.7560
Viroid	0.0006	0.0008	0.0009	0.0008	0.0008	0.0143
Other	99.7897	99.6800	99.6577	99.7140	99.7046	93.3404
Unidentified	0.1174	0.2245	0.2514	0.1970	0.2033	0.8894
MV5	Barcode 5	Barcode 6	Barcode 7	Barcode 8	Combined	Illumina
Virus	0.0987	0.0971	0.0929	0.0982	0.0968	3.6803
Viroid	0.0017	0.0006	0.0013	0.0011	0.0012	0.0150
Other	99.7469	99.6967	98.9887	99.6494	99.5091	95.4491
Unidentified	0.1527	0.2055	0.9171	0.2513	0.3929	0.8556
MV9	Barcode 9	Barcode 10	Barcode 11	Barcode 12	Combined	Illumina
Virus	0.2789	0.2888	0.2878	0.2516	0.2732	11.3271
Viroid	0.0041	0.0034	0.0037	0.0043	0.0040	0.0201
Other	99.5977	99.5597	99.5765	99.6124	99.5911	87.7048
Unidentified	0.1193	0.1481	0.1320	0.1317	0.1316	0.9479

*The percentage of identified contigs/reads using Blastn were categorised into four groups: virus, viroid, unidentified and other which refers to all other types of hits as well as host genome

Table 3 Kraken2 sequence classification per sample across illumina and Oxford nanopore technologies (ONT) platforms

Classification*	**Percentage of contigs / sequences classified (%)**					
MV4	Barcode 1	Barcode 2	Barcode 3	Barcode 4	Combined	Illumina
Unclassified	0.1330	0.5343	0.5821	0.6069	0.4846	5.3000
Bacteria	0.0008	0.0043	0.0047	0.0061	0.0042	0.1556
Fungi	0.0001	0.0006	0.0005	0.0005	0.0004	0.0076
Viruses	0.0373	0.0617	0.0576	0.0571	0.0546	8.0950
Viroids	0.0008	0.0007	0.0008	0.0008	0.0008	0.0260
Citrus sinensis	79.4408	84.3585	82.8263	85.9280	83.3546	79.6795
Other	20.3872	15.0399	16.5280	13.4006	16.1008	6.7362
MV5	Barcode 5	Barcode 6	Barcode 7	Barcode 8	Combined	Illumina
Unclassified	1.1332	0.7240	0.8987	0.7909	0.9030	5.2883
Bacteria	0.0781	0.0133	0.0111	0.0082	0.0287	0.1264
Fungi	0.0009	0.0009	0.0012	0.0009	0.0010	0.0099
Viruses	0.0561	0.0684	0.0572	0.0587	0.0590	7.5684
Viroids	0.0015	0.0006	0.0015	0.0010	0.0012	0.0192
Citrus sinensis	74.3720	84.8229	79.5019	78.4349	78.5117	82.0959
Other	24.3582	14.3699	19.5285	20.7054	20.4955	4.8920
MV9	Barcode 9	Barcode 10	Barcode 11	Barcode 12	Combined	Illumina
Unclassified	0.4898	0.4497	0.5296	0.4059	0.4603	6.1457
Bacteria	0.0335	0.0509	0.0381	0.0219	0.0340	0.4392
Fungi	0.0041	0.0049	0.0039	0.0058	0.0048	0.0167
Viruses	0.1501	0.1568	0.1539	0.1309	0.1456	8.6088
Viroids	0.0038	0.0031	0.0035	0.0039	0.0036	0.0263
Citrus sinensis	58.8556	59.0234	60.1470	56.8859	58.4527	72.9903
Other	40.4630	40.3112	39.1240	42.5457	40.8990	11.7730

*The four categories include viroids, specifically citrus viroids (species level), viruses representing citrus viruses (species level), bacteria (root, R1), fungi (kingdom), unclassified and *Citrus sinensis* (species level) as designated by Kraken2 taxonomic assignment. R1 is a taxonomic rank code indicating a taxon is between root and domain ranks. The category "other" represents all sequences not included in the six previously mentioned categories and includes sequences that were not classified to the level required for each category

Reference gene-associated read mapping

Illumina and ONT reads were mapped to *C. sinensis* reference genes to indicate the biological variation associated with each sample as well as to compare the expression profiles of genes across platforms (Additional file 2). Twelve genes were selected and the TPM count of each gene was utilised to compare the proportion of mapped reads to a specific gene (Additional file 4 and 5). Regarding the ONT data, there was little to no variation evident amongst the barcodes from a sample and even across samples, the overall portion of TPM associated with specific reference genes seemed to reiterate one another. This was also applicable to Illumina data where the TPM profiles of the three plant samples were congruent. There was, however, a clear distinction between the TPM profiles associated with ONT and Illumina data (Additional file 4). The genes with the highest average TPM in the ONT dataset were *GAPC1* closely followed by *UBC9* whereas the *TUB* gene had the highest average TPM for the Illumina data (Additional file 5). The significant difference in *TUB*'s TPM between platforms (p-value < 0.05) had a prominent impact on the differential TPM pattern observed (Additional file 6). Additionally, the stark contrast of the proportion of reads mapped to *SAND* (p-value < 0.05) and *DIM1* (p-value < 0.05) between ONT and Illumina also contributed to the visually distinct profiles observed (Additional file 4). The only genes that remained statistically similar between platforms were *UBC21* and *Unknown* (p-value > 0.05) (Additional file 6). Conversely, there was also no significant differences observed between the combined ONT dataset and the individual barcode datasets.

The gene coverage (span from the first to the last base of the reference sequence) for each reference gene per ONT sample is indicated in Additional file 7 as the coverage distribution of the associated barcodes from that sample. The gene coverage per Illumina sample is also indicated (diamond points). All reference genes in the Illumina data had an average gene coverage of more than 96%, except *UBC21* (88%), whereas the gene coverage of the combined ONT dataset (for each sample, square points) exceeded 93% for the majority of the reference genes. However, some genes had more varied gene coverage amongst barcode datasets, notably *BOX*, *SAND*, *PTB*, *Unknown* and *UPL7*. These genes were also associated with relatively low TPM counts (Additional file 5).

Pathogen-associated read mapping

Pathogen reference genomes for all three viruses and three viroids that correspond to the full pathogen

complement of the source plant were used in read mapping analyses to compare Illumina and ONT datasets. In order to achieve this, the TPM of each pathogen relative to the other pathogens in a specific sample was used to evaluate differences in relative abundance. Relevant CTV genotypes were also included, and the sum of all genotypes was used to determine the CTV component of each sample. Except for dataset MV5 Barcode 6, the expected pathogen profile was detected in all ONT barcode-specific datasets and Illumina datasets (Fig. 1). As it was with Blastn, Kraken2 and MiFi®, dataset MV5 Barcode 6 did not contain any reads that mapped to HSVd. In spite thereof, when comparing the overall TPM distribution of pathogens, it is evident that the viroid component was larger in ONT samples. Therefore, while the pathogen component was smaller in ONT datasets (as indicated by Kraken2 and Blastn), the relative viroid component within the pathogen component was larger compared to Illumina. Conversely, Illumina samples had a larger virus component within the pathogen component, but the distribution amongst viruses was noteworthy. CTV had the highest TPM count which accounted for an average of 85% of pathogen reads while in the ONT datasets, although it was still the pathogen with the highest TPM, it on average merely accounted for 38.9% of pathogen reads (Fig. 1). Therefore, CTLV and CiVA-RNA2 comprised a significantly larger portion of the ONT TPM profiles. Furthermore, statistically significant differences between the ONT and Illumina datasets (p-value < 0.05) were observed for all pathogens, except CiVA-RNA1 and HSVd (Additional file 8). As with the reference gene mapping, there were also no significant differences observed between the combined ONT dataset and the individual barcode datasets. Although the overarching TPM pattern and trends observed in barcode datasets within a specific plant were consistent, variation amongst barcode datasets was less pronounced in MV9 compared to MV4 and MV5 for all pathogens except CiVA-RNA1 where MV4 barcode datasets had the least amount of variation (lowest standard deviation).

Pathogen genome coverage

Virus and viroid read mapping also allowed the calculation of the fraction of each reference covered (genome coverage) per sample (Fig. 2). The genome coverage of CTLV, CDVd and CEVd was more than 99% across all ONT and Illumina datasets (Fig. 2). Contrary to this, ONT had more variability in HSVd genome coverage, particularly in the MV5 Barcode 6 dataset, while Illumina datasets all resulted in a consistent 100% coverage. CTV (RB) had the highest CTV genotype coverage and remained fairly consistent across Illumina and ONT datasets (Fig. 2). For the majority of remaining genotypes, Illumina sequencing conferred a higher genome coverage. When comparing the CiVA genomes, there was also a discrepancy between the performance of the platforms. While more than 96% coverage was obtained for RNA2 regardless of plant or platform, Illumina had an average coverage higher than 99% for RNA1 while ONT coverage had an expanded range (22% – 100%) across plant

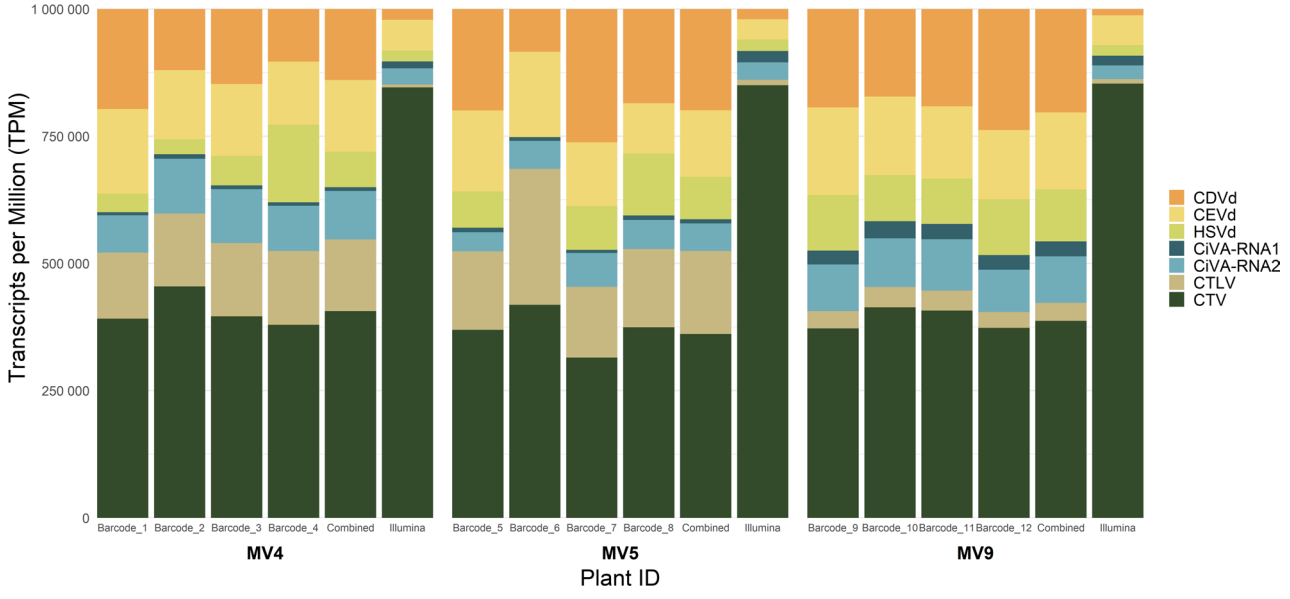

Fig. 1 Stacked column graph displaying the transcript per million (TPM) values for each pathogen reference pathogen genome across Oxford Nanopore Technologies (ONT) datasets and Illumina samples. Combined refers to the pooled data of all four barcode datasets of a specific sample. The pathogen reference genome abbreviations are listed on the right: citrus dwarfing viroid (CDVd), citrus exocortis viroid (CEVd), hop stunt viroid (HSVd), citrus virus A RNA1 and RNA2 (CiVA), citrus tatter leaf virus (CTLV) and citrus tristeza virus (CTV). Note that CTV as indicated here is depicted as the cumulative value of all genotype TPM counts included in the read mapping analysis: A18, T30, T3, T68, HA16-5, RB, VT and S1

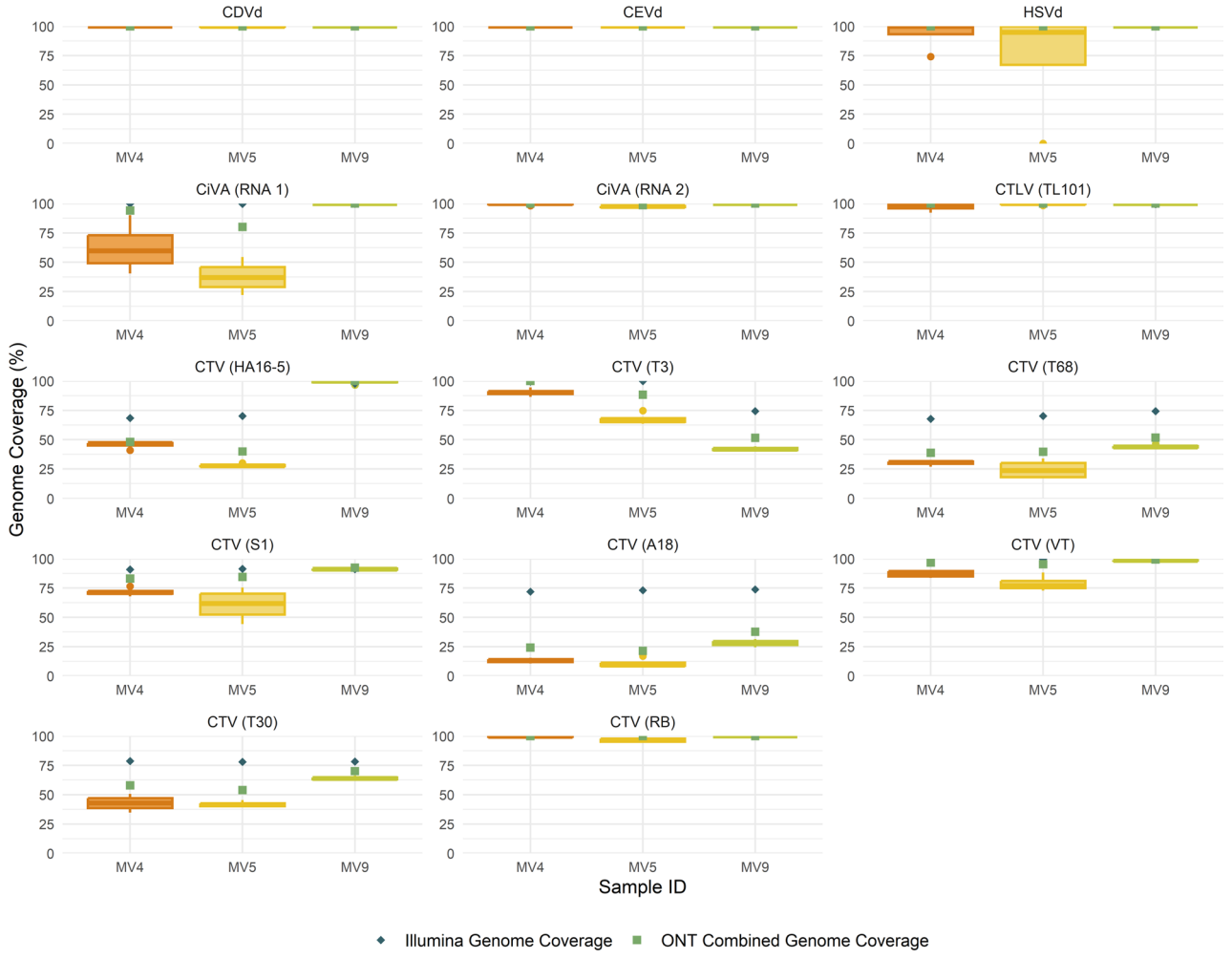

Fig. 2 Box and whisker plots of the genome coverage of each pathogen reference genome across samples. The boxplots represent the coverage distribution within a sample using Oxford Nanopore Technologies (ONT) sequencing (each sample has four barcode datasets). The square points are indicative of the combined dataset (pooled barcodes) genome coverage per sample. The diamond points show the coverage for each sample using Illumina sequencing. Pathogens included are citrus dwarfing viroid (CDVd), citrus exocortis viroid (CEVd), hop stunt viroid (HSVd), citrus virus A RNA1 and RNA2 (CiVA), citrus tatter leaf virus (CTLV) and citrus tristeza virus (CTV) genotypes A18, T30, T3, T68, HA16-5, RB, VT and S1

samples where only MV9 showed full genome coverage. Notwithstanding, it is important to note that the genome coverage offered by the dataset of the pooled (combined) barcodes was always equal to or, more often than not, higher than what was obtained with a singular barcode dataset. This was especially evident for CiVA-RNA1 in sample MV5 where a single barcode specific dataset on average managed 37.66% coverage and the combined barcode data improved coverage to 80.18% (Table 4).

Sequencing depth across pathogens

The sequencing depth of each pathogen was estimated as the number of reads that cover a single base of the reference sequence during read mapping. As expected, the sequencing depth obtained for each pathogen was higher for the combined ONT dataset in comparison to the average of a single barcode dataset per plant sample (Table 4). However, the Illumina sequencing depth was considerably higher than even the depth of the combined ONT dataset of all barcodes per plant sample. Despite this, more than 99% genome coverage could be achieved with ONT datasets with relatively low average depth for various pathogens including CDVd, CEVd, CiVA-RNA2, CTLV and CTV(RB) (Table 4). The depth distribution across pathogen genomes showed slight differences between sequencing platforms (Fig. 3, Additional file 9). The viroid depth profiles were similar for Illumina and ONT data although ONT depth distribution seemed to be more spread out across bases. Both Illumina and ONT had somewhat irregular coverage for CiVA-RNA1 across bases. Conversely, CiVA-RNA2 indicated a decrease in depth around the middle of the genome for both ONT and Illumina datasets. This decreased depth area is however larger (more bases affected) with

Table 4 Genome coverage (%) and average depth across plant samples

Pathogen	Accession	Average ONT barcode dataset		Combined ONT barcode dataset		Illumina dataset	
		Genome Coverage (%)	Average Depth	Genome Coverage (%)	Average Depth	Genome Coverage (%)	Average Depth
MV4							
CDVd	KY110718.1	100.00	5.88	100.00	32.05	100.00	747.45
CEVd	KY110721.1	100.00	7.96	100.00	36.90	100.00	2536.53
HSVd	KY110716.1	93.46	3.34	100.00	14.44	100.00	790.34
CiVA (RNA 1)	MT720885	62.52	3.72	94.20	5.57	100.00	637.20
CiVA (RNA 2)	MT720886	99.55	22.88	100.00	53.38	99.93	1513.50
CTLV (TL101)	MH108976.1	97.39	10.08	100.00	125.39	99.63	271.74
CTV (T30)	AF260651	42.72	2.41	57.89	17.38	78.63	2049.68
CTV (VT)	EU937519	87.08	8.42	96.59	28.06	99.90	5247.70
CTV (A18)	JQ798289.1	13.07	9.22	24.21	6.17	71.89	804.24
CTV (S1)	KU589212.1	71.56	9.34	83.00	13.16	90.78	2742.68
CTV (RB)	KU883265	99.55	34.17	100.00	89.58	100.00	18949.87
CTV (HA16-5)	KU883267.1	45.46	5.63	48.05	39.13	68.41	4635.73
CTV (T3)	MH051719	90.28	2.16	99.95	34.89	99.51	6198.44
CTV (T68)	MK033511	30.09	13.71	38.80	17.67	67.74	2504.27
MV5							
CDVd	KY110718.1	99.66	4.56	100.00	23.21	100.00	827.34
CEVd	KY110721.1	100.00	5.19	100.00	18.61	100.00	1817.52
HSVd	KY110716.1	96.47	2.97	100.00	9.78	100.00	933.32
CiVA (RNA 1)	MT720885	37.66	3.11	80.18	2.30	99.84	1201.38
CiVA (RNA 2)	MT720886	97.42	10.99	98.65	17.25	99.96	1830.28
CTLV (TL101)	MH108976.1	99.59	2.01	100.00	78.80	99.80	561.70
CTV (T30)	AF260651	41.74	3.38	53.88	14.14	78.02	2860.54
CTV (VT)	EU937519	78.76	5.82	95.30	17.14	99.81	9087.51
CTV (A18)	JQ798289.1	10.90	4.65	21.30	6.07	73.06	807.81
CTV (S1)	KU589212.1	60.76	4.42	84.43	8.95	91.34	4094.61
CTV (RB)	KU883265	96.70	19.78	100.00	42.51	100.00	19443.02
CTV (HA16-5)	KU883267.1	27.91	2.69	39.88	5.61	70.10	2192.69
CTV (T3)	MH051719	67.51	1.23	88.30	13.51	99.89	5466.91
CTV (T68)	MK033511	24.52	4.37	39.69	6.64	70.10	4168.93
MV9							
CDVd	KY110718.1	100.00	32.80	100.00	150.93	100.00	655.26
CEVd	KY110721.1	100.00	35.51	100.00	139.41	100.00	3548.47
HSVd	KY110716.1	100.00	26.31	100.00	92.09	100.00	1210.40
CiVA (RNA 1)	MT720885	100.00	42.03	100.00	82.52	99.87	1364.07
CiVA (RNA 2)	MT720886	100.00	87.20	100.00	240.97	100.00	1921.72
CTLV (TL101)	MH108976.1	100.00	9.15	100.00	149.13	99.58	608.39
CTV (T30)	AF260651	63.99	23.02	70.09	119.76	78.23	5721.07
CTV (VT)	EU937519	98.59	37.73	99.37	140.91	99.79	8402.00
CTV (A18)	JQ798289.1	27.99	34.85	37.70	78.13	73.77	1286.19
CTV (S1)	KU589212.1	91.12	23.21	92.37	165.86	91.14	8045.44
CTV (RB)	KU883265	100.00	37.28	100.00	348.81	100.00	27291.81
CTV (T3)	MH051719	42.07	18.98	51.59	29.86	74.33	2468.26
CTV (HA16-5)	KU883267.1	99.01	20.63	100.00	91.91	97.71	7007.04
CTV (T68)	MK033511	44.14	60.24	51.68	64.83	74.35	3797.14

Illumina sequencing. A striking difference in CTLV depth was observed because, unlike the Illumina datasets with even depth distribution, the ONT datasets had a gradual depth increase from the start to the middle of the genome followed by a sudden decrease and gradual incline again, that culminated in a relatively high depth at the end of the genome. The CTV distribution of depth obtained with ONT sequencing also varied with two major sequencing hotspots in the middle of the genome

Fig. 3 Bar graph distribution of the depth (number of reads) per base (position) of the pathogen reference genomes for MV4 across datasets. Blue bars show Illumina data. Orange bars represent a single Oxford Nanopore Technologies (ONT) barcode dataset, in this case Barcode 1. Green bars indicate the combined barcodes dataset. Only one CTV genotype (RB) is showed as a representative of all CTV genotypes. Pathogens included are citrus dwarfing viroid (CDVd), citrus exocortis viroid (CEVd), hop stunt viroid (HSVd), citrus virus A RNA1 and RNA2 (CiVA), citrus tatter leaf virus (CTLV) and citrus tristeza virus (CTV) genotype RB. For all other plant and barcode depth distribution graphs, see Additional file 9

not observed with Illumina although both platforms had an indication of increased depth at the 3' end (Fig. 3).

Virus and viroid read mapping consensus sequences

The ONT consensus sequences obtained for each pathogen from each sample's barcode specific dataset were evaluated for sequence identity compared to the consensus sequences obtained with Illumina datasets (Fig. 4). An overall average of 99% sequence similarity was obtained when comparing ONT consensus sequences to Illumina. In some datasets, 100% identity was shared between platforms for CEVd and CiVA-RNA2. There was also no considerable difference in identities obtained between the combined ONT datasets and the individual barcode datasets. For all three plants, CDVd had the lowest identity compared to Illumina sequences. Consensus sequences were also evaluated for the number of ambiguous bases present (Fig. 4). Both Illumina and ONT consensus sequences contained ambiguous bases for the majority of sequences obtained. CDVd sequences also had the overall highest percentage of ambiguous bases for ONT and Illumina. Notably, in some cases both the Illumina and ONT consensus sequences had ambiguous bases but when aligned, these bases were in the same position and had similar symbols leading to 100% sequence identity between platforms.

Platform cost comparison

The approximate cost of Illumina sequencing was four times less expensive than ONT sequencing for a single sample. The cost of Illumina sequencing using a service provider included all aspects of sequencing and a data delivery fee that applied once to a single order regardless of the number of samples within that order. The equivalent calculation for ONT sequencing was divided into the three main steps of sequencing preparation, and although library preparation was the costliest of these three steps, the largest expense of ONT was attributed to the flow cell. This was the consequence of the singular usage of flow cells, meaning that a new flow cell was used for each sample.

Sample	Pathogen	Percentage consensus sequence identity (%)					Percentage ambiguous bases in consensus sequence (%)					
		Barcode_1 - Illumina	Barcode_2 - Illumina	Barcode_3 - Illumina	Barcode_4 - Illumina	Combined - Illumina	Barcode_1	Barcode_2	Barcode_3	Barcode_4	Combined	Illumina
MV4	CDVd	96.59	98.29	97.62	97.27	97.95	2.03	3.38	1.68	4.41	4.39	4.08
	CEVd	100	100	99.46	99.19	100	0.27	0.27	0.27	0.27	0.27	0.27
	HSVd	99.33	99.55	99.67	99.67	99.66	0	0	0	0	0	0.33
	CiVA (RNA 1)	99.72	99.63	99.59	99.53	99.75	0.08	0.17	0.11	0.16	0.17	0
	CiVA (RNA 2)	99.82	99.85	99.89	99.85	99.85	0.18	0.15	0.11	0.15	0.15	0
	CTLV (TL101)	99.91	99.73	99.89	99.79	99.86	0.08	0.17	0.08	0.13	0.06	0.02
	CTV (VT)	98.23	98.51	98.1	98.03	98.49	0.75	0.89	0.57	0.67	1.62	1.57
	CTV (S1)	99.18	98.71	98.48	98.54	99.36	0.38	0.48	0.26	0.38	0.42	0.17
	CTV (RB)	98.96	98.95	99.06	99.01	98.85	1.09	1.17	1.08	1.06	1.34	0.21
	CTV (T3)	98.51	98.75	98.68	98.46	98.79	0.5	0.48	0.51	0.83	1.36	0.74
		Barcode_5 - Illumina	Barcode_6 - Illumina	Barcode_7 - Illumina	Barcode_8 - Illumina	Combined - Illumina	Barcode_5	Barcode_6	Barcode_7	Barcode_8	Combined	Illumina
MV5	CDVd	96.62	95.85	96.28	97.63	94.93	0	0	3.04	1.69	3.03	1.35
	CEVd	98.92	98.65	98.65	100	100	0	0.54	0	0.27	0.27	0.27
	HSVd	99.63	NA	99.34	99.67	99.34	0	NA	0.33	0	0.33	0.33
	CiVA (RNA 1)	99.33	99.93	99.58	99.67	99.66	0.1	0	0.04	0.05	0.13	0
	CiVA (RNA 2)	99.66	99.78	99.85	99.85	99.96	0.26	0.19	0.15	0.15	0.04	0
	CTLV (TL101)	99.6	99.48	99.86	99.86	99.88	0.08	0.11	0.11	0.09	0.08	0.03
	CTV (VT)	98.97	98.36	98.84	99.54	98.69	0.2	0.16	0.22	0.1	0.22	0.77
	CTV (S1)	98.48	97.36	98.83	98.91	99.34	0.14	0.02	0.05	0.17	0.17	0.09
	CTV (RB)	99.56	99.29	99.37	99.61	99.58	0.4	0.67	0.65	0.55	0.66	0.3
	CTV (T3)	99.24	99.09	99.28	99.39	99.48	0.26	0.08	0.09	0.17	0.34	0.28
		Barcode_9 - Illumina	Barcode_10 - Illumina	Barcode_11 - Illumina	Barcode_12 - Illumina	Combined - Illumina	Barcode_9	Barcode_10	Barcode_11	Barcode_12	Combined	Illumina
MV9	CDVd	95.25	94.56	95.27	94.93	94.56	4.39	2.37	4.39	4.04	4.75	3.04
	CEVd	99.73	99.73	99.73	99.73	99.73	0.54	0.54	0.54	0.54	0.54	0.27
	HSVd	99.67	100	100	99.67	99.67	0.33	0	0	0.33	0.33	0
	CiVA (RNA 1)	99.97	99.91	99.96	99.97	99.97	0.06	0.09	0.04	0.03	0.03	0
	CiVA (RNA 2)	100	100	100	100	100	0	0	0	0	0	0
	CTLV (TL101)	99.86	99.94	99.94	99.92	99.95	0.14	0.06	0.09	0.06	0.03	0
	CTV (VT)	99.64	99.47	99.43	99.66	99.7	0.07	0.07	0.05	0.04	0.03	0.03
	CTV (S1)	99.6	99.6	99.65	99.65	99.63	0.22	0.27	0.24	0.23	0.23	0.15
	CTV (RB)	99.56	99.25	99.33	99.25	99.56	0.73	1.04	0.97	0.94	0.73	0.3
	CTV (T3)	99.19	99.32	99.48	99.21	99.24	1.17	1.07	0.78	1.19	1.1	0.72

Lowest value (94.56%) Highest value (100%)

Lowest value (0%) 50th percentile (0.22%) Highest value (4.75%)

Fig. 4 Heatmaps of the percentage sequence identity of the Oxford Nanopore Technologies (ONT) read mapping consensus sequences compared to Illumina consensus sequences and the percentage of ambiguous bases present in ONT and Illumina consensus sequences. The sequence identity heatmap is calculated for the entire dataset ranging from the lowest value (light green) to 100% (dark green). Gaps and missing data (bases with no coverage) are not included in identity estimates and calculated identities only reflect the per base position identity comparison between platforms and not the completeness of consensus sequences. The ambiguous bases heatmap is calculated for the entire dataset where orange indicates the highest value and light yellow represents lowest value. The ambiguous bases accounted for include r, y, s, w, k, m, b, d, h and v. Combined refers to the pooled dataset of all four ONT barcode datasets of a specific sample. Pathogens included are citrus dwarfing viroid (CDVd), citrus exocortis viroid (CEVd), hop stunt viroid (HSVd), citrus virus A RNA1 and RNA2 (CiVA), citrus tatter leaf virus (CTLV) and citrus tristeza virus (CTV) genotypes T3, RB, VT and S1. These CTV genotypes were selected due to their relatively high average genome coverage (> 80%) across combined ONT datasets

Discussion

In this study, Illumina and ONT sequencing platforms were compared for their ability to reliably and accurately detect a panel of citrus viruses and viroids across three plant samples by using both reference-dependent and independent bioinformatics approaches. Within the context of this study, platform performance encapsulates both the library preparation and sequencing steps. For an extended comparison, the sequencing accuracy of each platform was also considered. Different ONT sequencing approaches might offer alternative advantages with accompanying limitations [5, 6, 8, 11]. cDNA-based ONT sequencing protocols have shown improved results for the sequencing of viruses and viroids compared to others, notably total RNA approaches [5, 6, 8, 11], and was, therefore, selected to compare sequencing platforms in this study. Importantly, the majority of previous studies of this nature utilised older ONT chemistries and, therefore, direct comparability was fairly limited.

Illumina sequencing was able to positively detect all pathogens included in this study across all samples in every bioinformatic methodology used regardless of whether reference sequences were used to identify pathogens. ONT nearly replicated this result, but the MV5 Barcode 6 dataset unfortunately tested negative for HSVd irrespective of the analysis strategy. Presumably, this is the consequence of flow cell variability which can influence the amount of sequencing data generated (Table 1) [47]. However, when considering more than one barcode dataset, or the entire combined dataset consisting of the four individual barcode datasets of MV5, detection of HSVd was possible. This emphasised

the importance of being aware of the limitations of ONT and highlighted the need for proper flow cell capacity dedicated to a sample (for instance more than a quarter of a flow cell) to overcome possible low pore availability and, consequently low data generation, which could potentially fail to detect low concentration pathogens like viroids. Within this context, flow cell capacity refers to the average amount of data that was generated per flow cell (Table 1) given the specific experimental conditions.

As mentioned, Illumina and the combined ONT datasets yielded the same diagnostic calls across all analyses and that extended to EDNA with MiFi®. This approach may prove valuable for quick and relatively cost-effective diagnostic calls [35]. Even with the inclusion of non-target viroid probes, no false positives were generated (Additional file 3). When considering the reference independent methods, both Blastn results and taxonomic sequence assignment yielded comparable results. This was evident although Kraken2 utilised trimmed reads for both the Illumina and ONT pipelines whereas the Blastn analyses used *de novo*-assembled contigs and unassembled trimmed reads, respectively. Given the relatively long read length and the difficulties commonly experienced with viral genome *de novo* assembly of ONT data, mostly attributed to insufficient read numbers and quality required for assemblers [6, 48], taxonomic classification tools like Kraken2, are viable alternatives to Blastn for viral pathogen identification. For both approaches, Illumina datasets contained a larger percentage of pathogen reads (Tables 2 and 3). While this was consistent with previous findings regarding relative viroid abundance, the prominent difference in virus amount between sequencing platforms was somewhat contradictory because previously ONT produced viral fractions comparable to Illumina [8, 11]. This can be a consequence of ONT library preparation, relative abundance calculation or the viruses included in this study considering the limited research currently available on ONT sequencing of citrus viruses [17, 18]. Additionally, previous studies utilised extracted RNA as input prior to library preparation and sequencing and not total nucleic acid as in this study which may also have contributed to the decrease of pathogen sequences obtained [8, 11]. Various difficulties have, however, been reported when utilising the ONT platform for viroid sequencing. Their circular genomes might lead to difficulties during library preparation with polyadenylation and their small genome size may be better suited for short-read sequencing as performed by Illumina platforms [8, 11]. The currently limited research available on viroid ONT sequencing emphasises the need for improvements and validation of all aspects of the sequencing pipeline including nucleic acid extraction, library preparation and bioinformatic diagnostic methodologies. In particular, library preparation poses a major variable considering the various kits currently available and being developed. Further enhancements to aid kit efficacy like the evaluation of alternative cDNA synthesis methods [11], in the case of the cDNA-based kits, could contribute towards the improved sequencing of certain pathogens like viroids.

The relative number of reads designated as host with Kraken2 was similar across platforms, despite the reference gene read composition varying considerably (Additional file 4). Evaluation of the reference gene expression profiles can provide valuable insights by functioning as an internal control for nucleic acid extraction efficiency and library preparation [1]. Reference genes can also be used to normalise pathogen relative read counts or TPM making direct comparison between samples possible. The twelve selected genes of this study were previously identified as reference genes in citrus [36] and assessing their relative expression profiles enabled comparison across samples and the identification of possible outliers. The distribution profiles within platforms were very consistent and along with the gene coverage (Additional file 7) associated with each gene, provided valuable insights as to which reference genes were more appropriate for a specific platform. In the ONT dataset, it was evident that reference genes with a lower TPM were associated with low gene coverage while Illumina samples retained a relatively high gene coverage across genes (> 87%) meaning that less genes were suitable reference genes for ONT sequencing. This may be due to sequencing biases, read length preferences or the impact of relative abundance (expression levels) of targets on accurate sequencing / identification with ONT and Illumina platforms, although further investigation is required. The consistency of the individual barcode-specific dataset's TPM counts (Additional file 4) of the same plant indicated that the workflow is reproducible and supports multiplexing on a single flow cell. Within a plant, all barcode datasets produced similar reference gene expression profiles which was also mirrored by the combined dataset.

In contrast to the consistency observed across the different barcode-specific ONT datasets of a plant sample for the reference gene expression profiles, there was a slight increase in variation observed in pathogen TPM ratios (Fig. 1). Comparing MV5 to MV4 and MV9, it was clear that more sequencing data lead to less variation between barcode-specific ONT datasets within a sample. Reference and pathogen composition profiles also supported the fact that reduced gene expression, or pathogens present in lower amounts were most adversely affected by inadequate data generation. Using more flow cell capacity, like combining multiple barcode datasets for one sample in this case, may circumvent potential issues associated with low data generation as seen with MV5. The most comparable results between ONT and

Illumina platforms were observed when considering the combined ONT dataset.

For all viroids, CiVA (especially RNA2), CTLV and the most abundant CTV genotypes (RB, VT, T3 and S1), ONT combined barcode datasets produced similar genome coverage to Illumina, while single barcode datasets varied more, most notably for HSVd, CTV (S1) and CiVA (RNA 1) (Fig. 2). This implied that, despite the low depth of ONT, even when different barcode-specific datasets were combined (Table 4), high and comparable genome coverage could be obtained due to the longer reads generated. Furthermore, ONT pathogen consensus sequences shared considerable identity (overall average of 99%) to that of Illumina consensus sequences (Fig. 4). CDVd had the lowest identities between platforms, but given the high number of ambiguous bases present in both ONT and Illumina sequences, this may reflect a cascading effect of the CDVd reference genome used for mapping, which might not be the best representation of the true CDVd genome within the plants. The elucidation of true sequence variation is further complicated by ambiguous bases that contribute towards sequence dissimilarity which can be a consequence of viral diversity or technical restraints (like sequencing quality). However, some instances where ambiguous bases aligned, resulting in complete similarity, supported the possibility that these bases may be the consequence of variability compared to the reference genome or the presence of mix variants in the plant.

The depth distribution along the length of the reference sequences showed similar trends for the majority of pathogens (Fig. 3). While these distributions were predominantly dispersed/even, CiVA-RNA-2 had a unique profile due to its genomic organisation. CiVA, a member of the genus *Coguvirus*, has a bipartite genome and the RNA2 segment has an ambisense coding strategy for ORF2a and ORF2b [49, 50]. These two ORFs are separated by a long intergenic AU-rich hairpin and given the position of this hairpin and the area with less depth seen in Fig. 3, it appears that this loop may have impeded effective sequencing of this region since this was observed for both ONT and Illumina. There was, however, some variability regarding the depth distribution between ONT and Illumina, particularly for CTLV (Fig. 3). CTLV is regarded as a variant of apple stem grooving virus (ASGV) a member of the genus *Capillovirus*, in the family *Betaflexiviridae*. It is a polyadenylated positive single-stranded RNA virus where open reading frame two (ORF2) located at the 3' end, is translated by sub-genomic RNAs [51, 52]. Present findings had reduced depth of CTLV at the 5' end and severe bias was seen at the 3' end in ONT datasets (Fig. 3). Similar genome depth distribution has been reported previously for MinION sequenced polyadenylated RNA viruses [53]. This was likely the consequence of the library preparation and sequencing strategy, which selected for polyadenylated fragments. This led to overrepresentation of certain parts of the genome, especially at the 3' end.

Overall, ONT was able to accurately detect the most prominent pathogen genomes in these plants similar to Illumina sequencing. Although the total pathogen reads component was smaller in ONT datasets, the reads within this component were more evenly distributed between viruses and viroids compared to Illumina sequencing where CTV had a dominant read count. This might point to ONT sequencing being more suited to detecting diverse viral pathogens within a sample although further investigation and validation are required. Regardless, results suggested that having more data can have a significant positive impact on the performance of ONT sequencing. In this study when at least two barcode datasets were combined, which is the equivalent of half a flow cell's capacity, accurate detection of all pathogens across all plants were enabled. The inclusion of more samples, as well as nucleic acid extraction technical replicates, can improve statistical power and characterisation of variation between plants across platforms.

The prospect of using barcoding kits for multiplexing could have certain sequencing cost implications. As it currently stands, within a South African research environment context, ONT sequencing exceeds the service provider cost of Illumina sequencing. If multiplexing is applied appropriately, the cost of ONT sequencing can be greatly reduced. For instance, within the context of this study, if four samples were run on the same flow cell with different barcodes, the cost per sample would be similar to the cost of Illumina sequencing provided that the flow cell capacity dedicated to each sample is sufficient. Currently, as observed in this study, a quarter of a flow cell's capacity might not produce sufficient data under certain conditions. With the pooling of libraries, there is also always a possible risk of cross-contamination [2, 11]. It is, therefore, imperative to reiterate the importance of being aware of the constraints of ONT to ensure the most appropriate application of this technology. In various situations, the advantages offered by ONT might outweigh the possible limitations. The ability to offer short turnaround times as well as real-time analysis of samples can greatly benefit routine diagnostic laboratories, therefore, contending with other platforms such as Illumina sequencing for the incorporation into plant quarantine and certification schemes.

Conclusions

This study encompasses a systematic comparison of ONT with Illumina sequencing based on their ability to effectively detect three citrus viruses and three viroids. Multiple pathogen detection workflows were applied

with both reference driven and independent methodologies. Given the present results, it is clear that the amount of data required to make a positive diagnostic call is not objective and varies between platforms. While less relative pathogen read abundance and depth were associated with ONT data compared to Illumina, the longer reads compensated for these factors which enabled accurate pathogen detection. Both Illumina and ONT platforms were able to identify all viruses and viroids irrespective of the bioinformatic approach. Sometimes single ONT barcode-specific datasets performed less optimally which was indicative of a poorer sequencing run most likely attributed to variability observed between flow cells (i.e. the number of active pores available at the beginning of a sequencing run). Additionally, combining ONT barcode-specific datasets from a sample, equivalent to using at least half of the flow cell's data generation capacity (i.e. the average amount of data that will be produced by the flow cell in a specific experimental setup), considerably increased the quality and reliability of results obtained and is, therefore, a recommended practice especially if the sample is of high importance. This study also showed that cost feasibility can be achieved through multiplexing of samples and along with other advantages like real-time analysis and short turnaround times offering scalable solutions, might be valuable for diagnostic laboratories.

The continued development and improvement of ONT creates a hopeful outlook on the future of this technology as constant improvements are being made. This, however encouraging, does cause some difficulties with incorporating ONT into routine standard practices. When considering or implementing new standard practices into existing programs, validation of all aspects is crucial. This is especially true for programs and efforts that involve the distribution of healthy plant material and plant disease management as various economically important agricultural industries could be adversely affected by the introduction and propagation of pathogen infected plant material. There has been a global increase in the uptake of HTS in plant certification and quarantine programs which further drives research and validation of available sequencing technologies and detection pipelines. The research presented here encapsulates direct contributions towards the application of HTS within a citrus industry context and provides an initial investigation of ONT as a possible detection tool.

Abbreviations

HTS	High-throughput sequencing
RNA	Ribonucleic acid
ONT	Oxford ONT Technologies
cDNA	Complementary DNA
PCR	Polymerase chain reaction
rRNA	Ribosomal RNA
rcf	Relative centrifugal force
CTV	Citrus tristeza virus
CTLV	Citrus tatter leaf virus
HSVd	Hop stunt viroid
CDVd	Citrus dwarfing viroid
CEVd	Citrus exocortis viroid
CiVA	Citrus virus A
cv.	Cultivar
RT-PCR	Reverse transcription polymerase chain reaction
ssRNA	Single stranded RNA
kb	Kilobase
bp	Base pair, CRAN: Comprehensive R Archive Network
EDNA	Electronic Diagnostic Nucleic Acid Analysis
E-probes	Electronic Probes
nt	Nucleotide
BWA	Burrows–Wheeler Alignment Tool
MAFFT	Multiple Alignment using Fast Fourier Transform
TPM	Transcripts per million
RPK	Read count per kilobase of sequence
Gb	Gigabase
CVd V	Citrus Viroid Five
CVd VI	Citrus Viroid Six
CVd VII	Citrus Viroid Seven
CRI	Citrus Research International
NCBI	National Center for Biotechnology Information
RdRp	RNA-dependent RNA polymerase
sgRNA	Sub-genomic-RNA
ORF	Open reading frame
GAPC1	Glyceraldehyde-3-phosphate dehydrogenase
EF-1a	Elongation factor 1-alpha
ACT2	Actin-2
FBOX	F-box family protein
UPL7	Ubiquitin-protein ligase 7
TUB	Beta-Tubulin
UBC21	Ubiquitin-conjugating enzyme 21
NA	Not Applicable
DIM1	Thioredoxin-like protein YLS8
SAND	Vacuolar fusion protein MON1 homolog
UBC9	Ubiquitin conjugating enzyme 9
PTB1	Polypyrimidine tract-binding protein 1

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12637-7.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Supplementary Material 7.

Supplementary Material 8.

Supplementary Material 9.

Acknowledgements

The authors would like to acknowledge Kobus Breytenbach from CRI for establishing all the plant material. Computations were performed using Stellenbosch University's HPC2: http://www.sun.ac.za/hpc. The bursary funding received from the Citrus Academy is hereby acknowledged.

Authors' contributions

MD performed the nucleic acid extraction, bioinformatic analyses, interpreted the results and drafted the manuscript; HJM interpreted the results and contributed to drafting the manuscript; RB obtained funding, designed the study, performed the ONT library construction and sequencing, conducted

bioinformatic analyses; interpreted the results and drafted the manuscript. All authors read and approved the final manuscript.

Funding
The authors thank Citrus Research International (CRI) for funding CRI project 1408.

Data availability
The datasets generated and analysed during the current study are available in the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI) as BioProject accession PRJNA1320971.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 September 2025 / Accepted: 5 February 2026
Published online: 11 February 2026

References

1. Bester R, Cook G, Breytenbach JHJ, Steyn C, De Bruyn R, Maree HJ. Towards the validation of high-throughput sequencing (HTS) for routine plant virus diagnostics: measurement of variation linked to HTS detection of citrus viruses and viroids. Virol J. 2021;18(1):61.
2. Maina S, Donovan NJ, Plett K, Bogema D, Rodoni BC. High-throughput sequencing for plant virology diagnostics and its potential in plant health certification. Front Hortic. 2024;3:1388028.
3. Singh A, Yasheshwar, Kaushik NK, Kala D, Nagraik R, Gupta S, et al. Conventional and cutting-edge advances in plant virus detection: emerging trends and techniques. 3 Biotech. 2025;15(4):100.
4. Lebas B, Adams I, Al Rwahnih M, Baeyen S, Bilodeau Guillaume J, Blouin AG, et al. Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: A step-by-step description. EPPO Bull. 2022;52(2):394–418.
5. Liefting LW, Waite DW, Thompson JR. Application of Oxford nanopore technology to plant virus detection. Viruses. 2021;13(8):1424.
6. Javaran VJ, Poursalavati A, Lemoyne P, Ste-Croix DT, Moffett P, Fall ML. Nano-Viromics: long-read sequencing of DsRNA for plant virus and viroid rapid detection. Front Microbiol. 2023;14:1192781.
7. Talon M, Caruso M, Gmitter FG jr. The genus citrus. Cambridge, UK: Woodhead Publishing; 2020.
8. Pecman A, Adams I, Gutierrez-Aguirre I, Fox A, Boonham N, Ravnikar M, et al. Systematic comparison of nanopore and illumina sequencing for the detection of plant viruses and viroids using total RNA sequencing approach. Front Microbiol. 2022;13:883921.
9. Maree HJ, Fox A, Al Rwahnih M, Boonham N, Candresse T. Application of HTS for routine plant virus diagnostics: state of the Art and challenges. Front Plant Sci. 2018;9:1082.
10. Villamor D, Ho T, Al Rwahnih M, Martin R, Tzanetakis I. High throughput sequencing for plant virus detection and discovery. Phytopathology. 2019;109(5):716–25.
11. Mosquera-Yuqui F, Ramos-Lopez D, Hu X, Yang Y, Mendoza JL, Asare E, et al. A comparative template-switching cDNA approach for HTS-based multiplex detection of three viruses and one viroid commonly found in Apple trees. Sci Rep. 2025;15(1):1657.
12. Bester R, Steyn C, Breytenbach JHJ, de Bruyn R, Cook G, Maree HJ. Reproducibility and sensitivity of high-throughput sequencing (HTS)-based detection of citrus tristeza virus and three citrus viroids. Plants. 2022;11(15):1939.
13. Branton D, Deamer D. The development of nanopore sequencing. In: Nanopore sequencing. London: World Scientific; 2019. p. 1–16.
14. Van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34(9):666–81.
15. Lee HJ, Kim SM, Jeong RD. Analysis of wheat Virome in Korea using illumina and Oxford nanopore sequencing platforms. Plants. 2023;12(12):2374.
16. Oxford Nanopore Technologies. Product update: products scheduled to be discontinued [Available from: https://nanoporetech.com/es/products/prepare/discontinued-kits.
17. Johnson PZ, Needham JM, Lim NK, Simon AE. Direct nanopore RNA sequencing of umbra-like virus-infected plants reveals long non-coding RNAs, specific cleavage sites, D-RNAs, foldback RNAs, and temporal- and tissue-specific profiles. NAR Genomics Bioinf. 2024;6(3):lqae104.
18. Kim HJ, Cho IS, Choi SR, Jeong RD. Identification of an isolate of citrus Tristeza virus by nanopore sequencing in Korea and development of a CRISPR/Cas12a-based assay for rapid visual detection of the virus. Phytopathology. 2024;114(6):1421–8.
19. Hadidi A, Czosnek HH, Kalantidis K, Palukaitis P. Viroids and satellites and their vector interactions. Viruses. 2024;16(10):1598.
20. Bester R, Cook G, Maree HJ. Citrus Tristeza virus genotype detection using high-throughput sequencing. Viruses. 2021;13(2):168.
21. Bester R, Karaan M, Cook G, Maree HJ. First report of citrus virus A in citrus in South Africa. J Citrus Pathol. 2021;8(1):1–6.
22. Arruabarrena A, Benitez-Galeano MJ, Giambiasi M, Bertalmio A, Colina R, Hernandez-Rodriguez L. Application of a simple and affordable protocol for isolating plant total nucleic acids for RNA and DNA virus detection. J Virol Methods. 2016;237:14–7.
23. Onate-Sanchez L, Vicente-Carbajosa J. DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC Res Notes. 2008;1:93.
24. Oxford Nanopore Technologies. 3' polyadenylation of RNA transcripts using E. coli poly(A) polymerase (PAP): Oxford Nanopore Technologies. 2023 [updated 07/11/2023. Available from: https://nanoporetech.com/document/extraction-method/3-poly-rna-ecoli-pap.
25. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinf. 2010.
26. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11(10):e0163962.
27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
28. Bester R, Maree HJ. Validation of High-Throughput sequencing (HTS) for routine detection of citrus viruses and viroids. Viral Metagenomics: Methods and Protocols: Springer; 2023. pp. 199–219.
29. De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9.
30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
31. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. Research in Computational Molecular Biology: 17th Annual International Conference RECOMB 2013 Proceedings. Springer Berlin Heidelberg; 2013. p. 158–70.
32. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20(1):257.
33. R Core Team. R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing. 2013 [Available from: https://www.R-project.org/.
34. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-; 2016.
35. Nascimento D, Bodaghi S, Wang H, Ribeiro-Junior M, Campos R, Dang T, et al. Development and validation of a suite of E-Probes for electronic diagnostic nucleic acid analysis (EDNA) for 20 Graft-Transmissible pathogens of citrus using MiFi and blind ring testing among novice users. PhytoFrontiers™. 2025;5(2):243–53.
36. Mafra V, Kubo KS, Alves-Ferreira M, Ribeiro-Alves M, Stuart RM, Boava LP, et al. Reference genes for accurate transcript normalization in citrus genotypes under different experimental conditions. PLoS ONE. 2012;7(2):e31263.
37. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
38. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
39. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37(23):4572–4.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.

41. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of samtools and BCFtools. Gigascience. 2021;10(2):giab008.
42. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
43. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583–621.
44. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. Wiley; 2013.
45. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc. 1961;56(293):52–64.
46. Dunn OJ. Multiple comparisons using rank sums. Technometrics. 1964;6(3):241–52.
47. Otron DH, Filloux D, Brousse A, Hoareau M, Fenelon B, Hoareau C, et al. Improvement of nanopore sequencing provides access to high quality genomic data for multi-component CRESS-DNA plant viruses. Virol J. 2025;22(1):78.
48. Amoia SS, Minafra A, Nicoloso V, Loconsole G, Chiumenti M. A new Jasmine virus C isolate identified by nanopore sequencing is associated to yellow mosaic symptoms of *Jasminum officinale* in Italy. Plants 2022;11(3):309.
49. Navarro B, Minutolo M, De Stradis A, Palmisano F, Alioto D, Di Serio F. The first phlebo-like virus infecting plants: a case study on the adaptation of negative-stranded RNA viruses to new hosts. Mol Plant Pathol. 2018;19(5):1075–89.
50. Navarro B, Zicca S, Minutolo M, Saponari M, Alioto D, Di Serio F. A negative-stranded RNA virus infecting citrus trees: the second member of a new genus within the order *Bunyavirales*. Front Microbiol. 2018;9:2340.
51. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the international committee on taxonomy of viruses (ICTV). Nucleic Acids Res. 2018;46(D1):D708–17.
52. Vats G, Sharma V, Noorani S, Rani A, Kaushik N, Kaushik A, et al. Apple stem grooving capillovirus: pliant pathogen and its potential as a tool in functional genomics and effective disease management. Archives Phytopathol Plant Prot. 2024;57(4):261–95.
53. Wongsurawat T, Jenjaroenpun P, Taylor MK, Lee J, Tolardo AL, Parvathareddy J, et al. Rapid sequencing of multiple RNA viruses in their native form. Front Microbiol. 2019;10:260.

Publisher's note