Nucleic Acids Research, 2025, **53**, gkaf221
https://doi.org/10.1093/nar/gkaf221
Computational Biology

A method for in-depth analysis of circular DNA virus populations by unambiguously profiling the low abundant virus variants and partial genomic components

Victor Golyaev [1,2,*], Sam Dierickx[3], Koen Deforche[3], Wim Dumon[3], Hervé Vanderschuren [1,2,4,*]

[1]Tropical Crop Improvement Laboratory, Crop Biotechnics, Department of Biosystems, KU Leuven, Leuven 3001, Belgium
[2]KU Leuven Plant Institute (LPI), KU Leuven, Leuven 3001, Belgium
[3]Emweb BV, Herent 3020, Belgium
[4]Plant Genetics and Rhizospheric Processes Laboratory, Gembloux Agro BioTech, University of Liège, Gembloux 5030, Belgium

[*]To whom correspondence should be addressed. Tel: +32 81 62 25 71; Email: herve.vanderschuren@kuleuven.be
Correspondence may also be addressed to Victor Golyaev. Tel: +32 16 19 44 04; Email: victor.golyaev@kuleuven.be

Abstract

Severe epidemic outbreaks of diseases associated with newly emerging strains of single-stranded DNA (ssDNA) viruses have led to serious economic losses of numerous important food crops. While the current mitigation strategies are mostly relying on the deployment of genetic resistance in crop varieties, the constantly evolving virus populations have the potential to rapidly break virus resistance. Therefore, the development of diagnostic tools enabling early detection of virus variants associated with hypervirulence and/or expansion to new host species is urgently needed as an effective mitigation solution. Here, we introduce a novel approach by designing a pipeline that allows accurately identifying and characterizing the full-length sequence variants of viral circular DNA genomes utilizing Nanopore sequencing technology and the bioinformatics tool Genome Detective. We demonstrate that the pipeline is suitable to provide an accurate and in-depth analysis of monopartite *Tomato yellow leaf curl Sardinia virus* (TYLCSV) and multipartite *Banana bunchy top virus* (BBTV) ssDNA virus populations resulting in the profiling of high- and low-frequency virus variants with ≥1% relative abundance. The approach also enabled the unambiguous detection and characterization of four TYLCSV partial genomic sequences as well as several partial genomic sequences for each BBTV genomic component not previously reported and accumulating during infection.

Graphical abstract

Introduction

The emergence of second-generation high-throughput sequencing (SGS) technologies, such as Illumina, has revolutionized the field of virology by facilitating the discovery of novel viral species, enabling the rapid and cost-effective sequencing of entire virus genomes, including those of DNA viruses [1]. Because the Illumina platform is based on the use of relatively short read sequencing (SRS; 50- to 300-bp read lengths) and subsequent assembly of these reads into so-called consensus sequences, it has limitations for the in-depth profiling of

Received: August 3, 2024. **Revised:** February 19, 2025. **Editorial Decision:** March 9, 2025. **Accepted:** March 11, 2025

full-length virus genomes [2,3]. Viruses are often present in complex populations of highly similar sequences, known as viral quasispecies [4], and accurate assembly of these genetic variants using short reads is challenging [5–7]. Given this constraint, the third-generation sequencing (TGS) platforms, such as Oxford Nanopore (ONT) and Pacific Biosciences (PacBio), represent a suitable alternative for reconstructing full-length haplotypes of viral quasispecies with long reads. The main issue with TGS-based haplotype reconstruction is relatively low TGS read accuracy compared with SGS, especially within homopolymeric regions [8,9]. However, the latest improvements in PacBio and Nanopore sequencing protocols enable production of high-quality reads with a 99.9% accuracy and less or no bias for deletions/indels in homopolymers [10,11]. In this context, ONT technology, which offers a higher overall data output than PacBio and a lower cost per Gb [12,9], emerges as the optimal choice for reconstructing full-length virus sequences without assembly. This approach contrasts with the genomic assembly pipelines commonly used for plant virome analysis, which may introduce bias in the detection of high-frequency mutations and do not allow for the reconstruction of virus genomic variants and low-abundance partial genomic components.

Geminiviridae and *Nanoviridae* are the two most important single-stranded DNA (ssDNA) virus families infecting plants [13,14]. *Geminiviridae* consists of mono- and bipartite ssDNA viruses while *Nanoviridae* comprises multipartite viruses (six–eight components). The circular DNA genomes from both families do not encode polymerases and the host plant machinery is recruited to support virus replication via rolling-circle replication and recombination-dependent replication mechanisms [15,16]. Although the error rate of DNA virus replication mechanisms appears lower than that of RNA viruses which are replicated by low-fidelity RNA polymerases, multiple observations confirm that genetic diversity of plant ssDNA viruses might be similar to that reported for plant RNA viruses [17–19]. This diversity, represented by a dynamic swarm of viral genetic variants, is a result of the continuous evolution of the virus within the host plant, effectively as a quasispecies.

The primary factors influencing the composition of ssDNA virus populations are the processes of specific selection or genetic drift. Both the environment in which a virus population proliferates, and the host plant antiviral responses can exert selective pressure [19,20]. Importantly the diversification of mutant spectra within the same strain under varying selection pressures in different environments can result in a distinct mutational composition, while maintaining the consensus sequence unchanged [21–23]. Thus, the identification and in-depth profiling of virus genetic variants and associated mutations are essential for studying evolution of plant ssDNA virus populations and their adaptation to a new host or changing environmental conditions.

The main challenge of virus genetic variants profiling using TGS data is to distinguish "true" mutation from sequencing errors [6]. The currently available long-read assembly algorithms are either not optimized for analysis of viral populations with highly similar haplotypes [24–28] or they require high-quality short reads as an input rather than error-prone TGS data [29–31]. While several attempts have been made to design bioinformatics tools allowing virus haplotype reconstruction from low-accuracy TGS reads, their sensitivity does not allow detecting haplotypes with low abundance (1%) and diversity (<0.3%) [6,32,33]. To address these constraints, our pipeline generates and profiles high-quality full genome sequences of circular ssDNA virus variants with single nucleotide difference and 1% relative abundance in three steps. First, viral circular DNAs enriched from the total DNA sample by rolling circle amplification (RCA) are sequenced with the latest R10.4.1 Nanopore sequencing kit. The long RCA viral reads are then deconcatenated by TideHunter (TH) tool [34] to produce highly accurate intact viral full-genome sequences. Finally, viral sequences are clustered based on their similarity, and random errors corrected and quantified by Genome Detective online tool [35]. The complete bioinformatics pipeline starting from deconcatenation step is implemented into Genome Detective web interface that allows users to directly upload the raw Nanopore data and profile ssDNA virus variants in the samples.

Materials and methods

Infected plant materials

Zucchini plant samples infected with *Tomato yellow leaf curl-Sardinia virus* (TYLCSV) viral clone (GenBank: Z25751.1) by agroinfiltration and propagated for several months at the UMA-CSIC greenhouse facility (Estación experimental IHSM La Mayora, UMA-CSIC, Malaga, Spain). Banana plant samples infected with *Banana bunchy top virus* (BBTV) were collected in the fields of Democratic Republic of the Congo (DRC Congo) and North (Nghe An) and South (Lam Dong) Vietnam.

DNA extraction and virus enrichment

Leaf tissue samples (50 mg) were homogenized in 2 ml Eppendorf tube by grinding in 1 ml of CTAB buffer Cetyltrimethylammonium bromide (CTAB) 2%, Polyvinylpyrrolidone (PVP) 2%, NaCl 2 M, Tris 100 mM (pH = 8), ethylenediaminetetraacetic acid 25 mM (pH = 8)] using a sterile pestle. The homogenate was mixed twice with an equal volume of chloroform:isoamyl alcohol (24:1) and centrifuged at 10k rpm followed by collection of the aqueous phase. The total DNA was precipitated with 0.6 volume of isopropanol and 0.1 volume of NaOAc (3 M, pH 5.2) at −20°C for 1 h, recovered by centrifugation at 12k rpm for 20 min at 4°C, washed twice with 70% ethanol, and air-dried in a bio-cabinet.

High-quality viral DNA for sequencing was prepared with circular DNA enrichment method [36]. Briefly, extracted total DNA (20–50 ng) was enriched for circular DNAs by randomly primed circular DNA amplification (RCA) at 30°C for 18 h in a reaction containing 1× Phi29 DNA polymerase buffer, 1 mM dNTPs, 50 μM exo-resistant random primer, 0.02 U inorganic pyrophosphatase, and 10 U Phi29 DNA polymerase. The amplification products were purified by precipitation with 3 M sodium acetate, 20 mg/ml molecular biology-grade glycogen, and 100% ethanol, debranched via non-primed Phi29 amplification with 5 U of Phi29 DNA polymerase at 30°C for 2 h, and linearized by treatment with 50 U S1 endonuclease at 37°C for 30 min. Debranched DNA was purified with 3 M sodium acetate, 20 mg/ml molecular biology-grade glycogen, and 100% ethanol.

Library preparation and Nanopore sequencing

TYLCSV viral DNA was sequenced using an ONT MinION Flow Cell (R10.4.1). First, debranched double-stranded DNA

(dsDNA) sample (∼1000 ng) was mixed with the components of NEBNext FFPE DNA Repair and NEBNext Ultra II End Repair/dA-Tailing modules and incubated at 20°C for 15 min followed by enzymes inactivation at 65°C for 5 min. Subsequently, DNA was cleaned using the AMPure XP Beads followed by adapter ligation with Nanopore Ligation Sequencing Kit v14 according to the manufacturer's protocols.

Multiplex sequencing of BBTV viral DNA samples was performed with Nanopore Native Barcoding Kit v14 according to the manufacturer's protocol. Briefly, sequencing barcodes were attached to the repaired and end-prepped BBTV DNAs (∼1000 ng). The samples were pooled and then cleaned with AMPure XP Beads kit. Sequencing adapters were ligated to the barcoded samples with Nanopore Ligation Sequencing Kit v14 according to the manufacturer's protocol.

Prepared libraries (∼200 ng) were loaded into the Nanopore flow-cells (R10.4.1) according to the manufacturer's instructions and sequenced for 48–72 h using MinKNOW software.

Reads basecalling was performed with Guppy (v6.5.7) using super accuracy algorithm (dna_r10.4.1_e8.2_400bps_sup). Basecalled reads were demultiplexed, trimmed of adapters by Porechop (v0.2.4), and their quality was assessed by NanoStat (v1.6.0).

Illumina sequencing and data analysis

RCA-amplified and debranched TYLCSV and BBTV viral DNA samples were sequenced with whole genome 150-bp paired-end SGS according to manufacturer's protocol (BGI, Hong Kong, China).

Raw data with adapter sequences or low-quality sequences were processed by SOAPnuke software developed by BGI [37] and mapped to the TYLCSV and BBTV virus reference sequences by Minimap2 (v2.26) [38]. Mapping files were visualized and analyzed by Integrative Genomics Viewer (IGV) [39].

Genome detective algorithm description

The algorithm for detecting and characterizing full-length virus sequences using ONT was implemented into the online software platform Genome Detective [35] and is freely available for academic use. The input FASTQ files containing ONT raw or duplex read data after adapter trimming and demultiplexing are processed in several steps. First, TH tool is used to create high-quality consensus reads from the multiple virus sequence copies within each RCA-generated read. TH quantifies the number of viral copies in the raw/duplex reads and generates the consensus sequence. By default, the reads with a minimum of three copies are processed (Fig. 1B). The FASTQ files with the consensus sequences, generated by TH, are used for the similarity search step. At this step, non-viral reads are filtered based on nucleotide similarity with the NCBI RefSeq virus reference database using Kraken2 [40] and protein similarity with the SwissProt Uniref90 database using Diamond [41]. Selected reads are further assigned to one or more candidate viral reference genomes using BLASTN against the NCBI RefSeq virus database and sub-sampled in batches until the depth of coverage of 100 is obtained. The similarity level of the reads with the selected reference candidate is evaluated by global alignment scores considering simultaneous nucleotide and amino acid similarity using annotated global aligner scores [42] to disambiguate closely related candidates. Each read is subsequently assigned to the best candidate reference and assembled into a draft consensus sequence. The original consensus reads are then aligned to all draft consensus sequences using Mimimap2. The generated alignment BAM file is filtered to ensure that each reference has a single alignment by removing non-primary alignments using SAMtools [43]. The generated alignment is processed with BCFtools [43] to generate a VCF file that captures the genetic variability at each covered position.

At the last step, for each annotated virus group, the assigned reads are analysed for variations to reconstruct virus haplotypes. This is implemented using a sub-clustering algorithm to detect identical reads assigned to the particular variant, taking into account their circular nature and possible sequencing errors. The assigned reads are aligned with the starting position of the reference sequence. This requires the mapping of all assigned reads to the reference, which generates multiple partial linear alignments. All linear alignments of a single read are grouped to obtain the correct position of the read corresponding to the reference (Supplementary Fig. S1). This is implemented by using Minimap2 to perform the alignment, Samtools to group all alignments of a single read, and HTSLib library [44] to extract the cigar representation of every alignment from the BAM file and to create one complete linear alignment starting from the first nucleotide of the reference sequence.

To improve read clustering, sequencing errors are corrected. For that, the genetic variation in the reads at each genomic position is compared with the theoretical variation expected from the average error rate of the sequencing technology, which was set to 1% based on ONT simplex raw data (Q20) accuracy. To account for the finite number of reads being used in the multiple sequence alignment, the actual threshold for a minimum count of an allele at each position is calculated using a binomial distribution. More specifically, this threshold is calculated for every position based on the observed number of mutations (k), total coverage (n) of that position in the genome, and the average error rate (p) of the sequencing technology used, as

$$F\left(k; n, p\right) = \Pr\left(X \leq k\right) = \sum_{i=0}^{|k|} \binom{n}{i} p^i (1-p)^{n-i}.$$

The variant is kept at the given position only when Pr ($X < k$) is above 95%; otherwise, it is corrected.

Gaps are considered in the same way as a substitution. If there is only one significantly abundant nucleotide, it is used to correct the errors, while in case there are several nucleotides for the same position, the errors are corrected with the IUPAC ambiguity codes. The reference nucleotide is used if none of the bases are significant. Single gaps are corrected when the gap is not marked as significant. Other insertions or deletions are not evaluated.

After correction, the software clusters identical sequences and reports the variant when the threshold of minimum 10 sequences is reached. The variants with <10 sequences are filtered out based on the regular presence of indels in homopolymeric regions. The abundance of the detected variant is quantified and it is reported in the Genome Detective results as an individual annotated virus sequence.

Variant sequence analysis

Both Nanopore raw data and high-quality duplex reads, selected by duplex_tools (v0.3.3) with super accuracy algorithm (dna_r10.4.1_e8.2_400bps_sup), were uploaded into

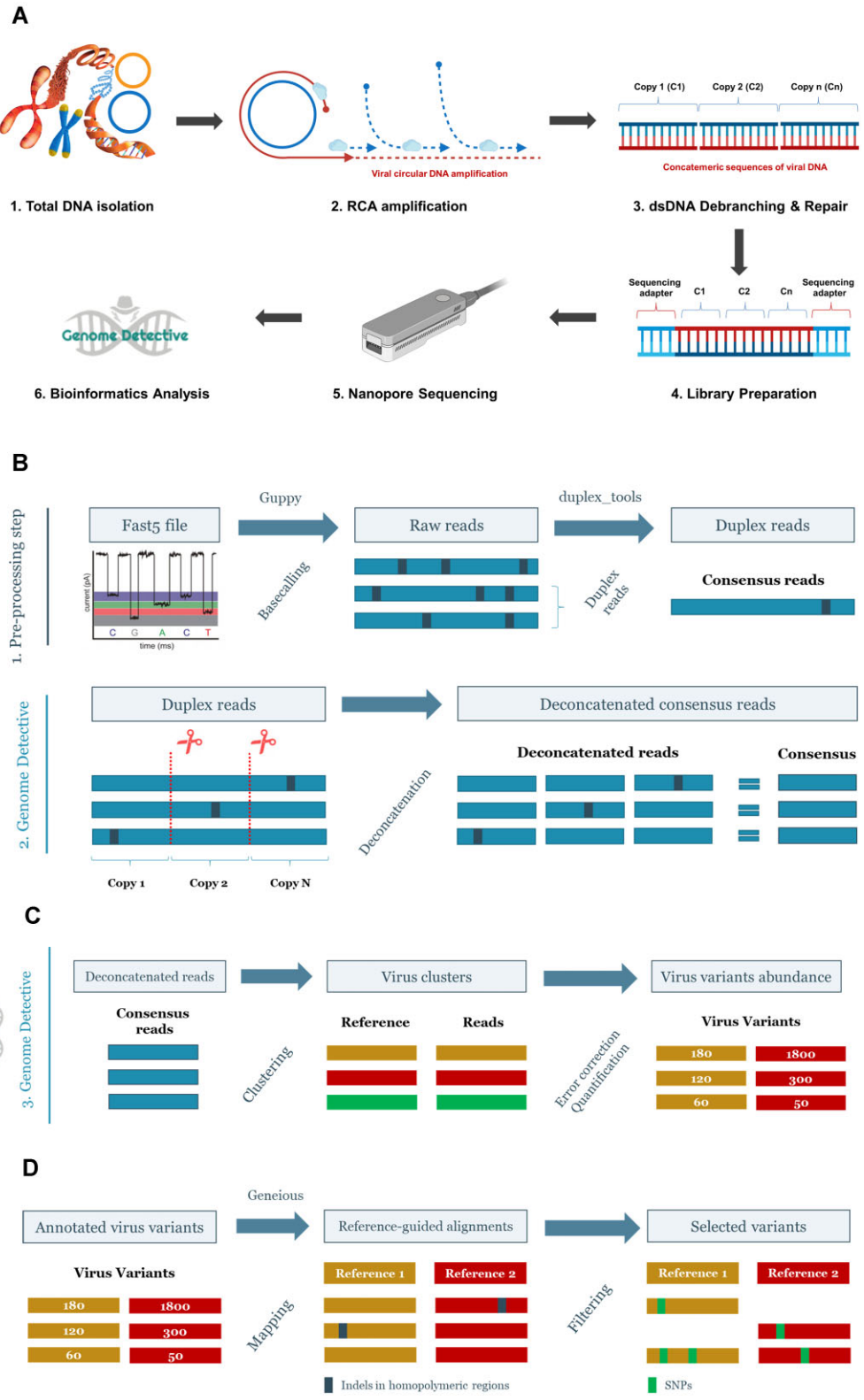


Figure 1. The pipeline for identification and analysis of circular ssDNA virus populations. (**A**) The pipeline workflow. (1) CTAB-based total DNA extraction. (2) Circular virus DNA enrichment via randomly primed RCA. (3) Enzymatic debranching of the RCA product with S1 nuclease. (4) dsDNA repair and adapter ligation. (5) Loading into the Nanopore flow-cell. (6) Detection of virus variants by Genome Detective and their abundance quantification. (**B**) Nanopore data analysis workflow. (1) The raw sequencing data are obtained by the neural network Guppy basecaller software followed by high-quality duplex reads selection using duplex_tools. (2) The duplex reads are uploaded into Genome Detective for deconcatenation using TH tool with error-corrected monomeric consensus sequences as an output. (**C**) Genome Detective post-processing workflow. (3) The deconcatenated sequences are clustered by mapping against RefSeq database, error-corrected, and their abundance is quantified by Genome Detective algorithm. (**D**) Virus variants analysis. Viral sequences annotated by Genome Detective were mapped against TYLCSV or BBTV reference sequences using Geneious software (v2023.2.1). Subsequently, virus sequences with deletions and/or insertions in homopolymeric regions were filtered out, and potential virus variants were used for SNP analysis with Geneious.

Genome Detective followed by deconcatenation with TH tool (v1.5.4) using default parameters (minimum of three copies of tandem repeats), clustered by mapping against RefSeq database, error-corrected, and their abundance was quantified.

For TYLCSV analysis, detected viral sequences were downloaded from Genome Detective and mapped against the TYLCSV reference (GenBank: Z25751.1) using Geneious software (v2023.2.1). For BBTV analysis, the most abundant BBTV viral variant sequences were used as reference sequences. Subsequently, TYLCSV and BBTV virus sequences with insertions in homopolymeric regions were filtered out, while sequences with deletions were re-analyzed and gaps were discarded, according to ICTV recommendations [45]. Potential virus variants were used for single nucleotide polymorphism (SNP) analysis with Geneious (Fig. 1D). Detected variants, their abundance, and SNPs were visualized by R packages ggplot2 and ggballoonplot.

To estimate the impact of the detected SNPs on the viral consensus sequences, TYLCSV and BBTV full-genome consensuses were reconstructed by Genome Detective using default viral metagenomic pipeline [35].

Results

In-depth profiling of monopartite DNA virus populations allows unambiguous identification of variants and partial viral genomes

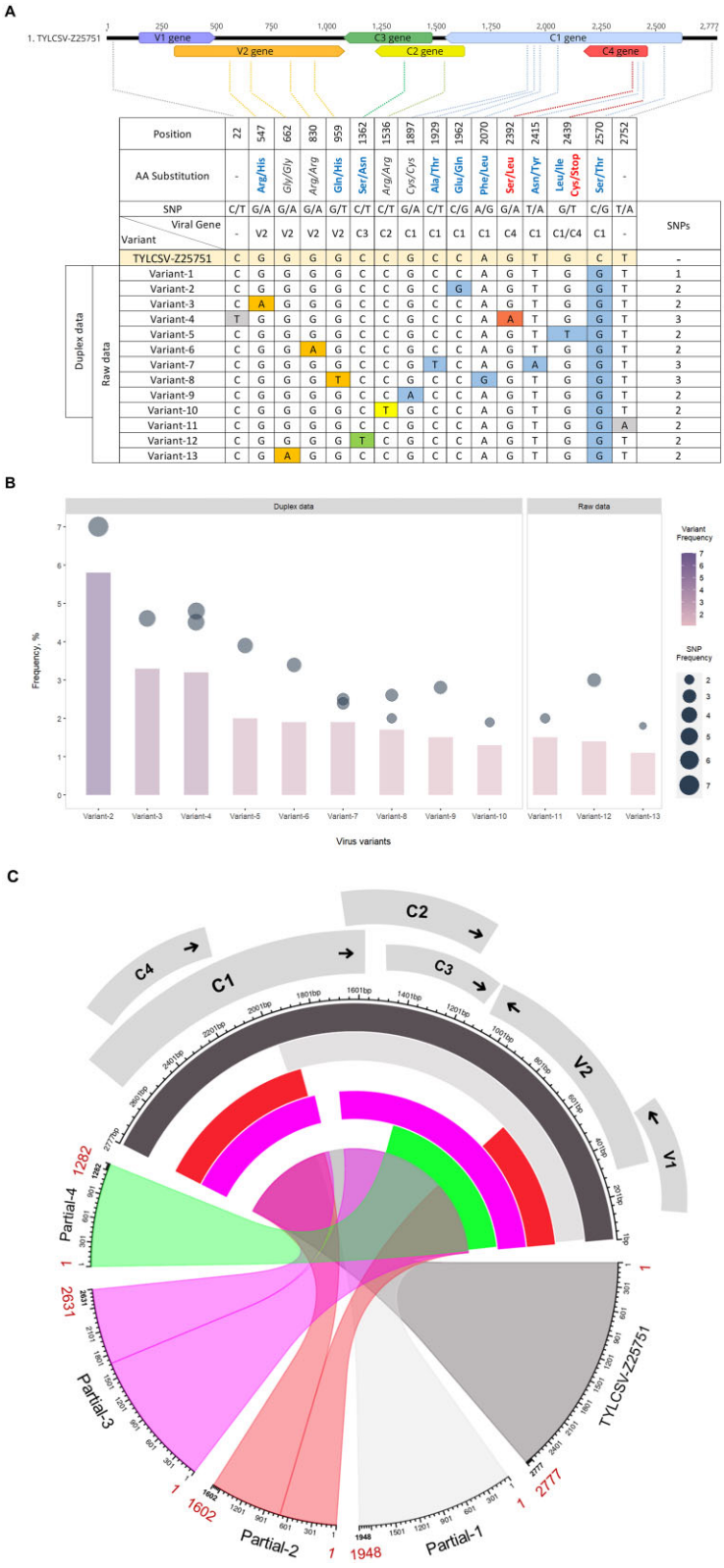
To assess the accuracy and sensitivity of our pipeline for ssDNA virus variant reconstruction, we profiled TYLCSV virus populations from zucchini plant samples infected with a TYLCSV infectious clone. Following sample preparation and Nanopore sequencing (see the "Materials and methods" section, Fig. 1A), high-quality duplex reads were obtained from raw sequencing data (Fig. 1B) and the resulting datasets were processed with Genome Detective online tool (see the "Materials and methods" section, Fig. 1C).

The reconstructed sequences were analyzed by mapping to the reference sequence of the TYLCSV infectious clone (GenBank: Z25751.1), and potential virus variants with no deletions and/or insertions in homopolymeric regions were selected (see the "Materials and methods" section, Fig. 1D). While 35% of the detected virus sequences had no issues with the correct reconstruction of homopolymeric regions, the other sequences were filtered out after careful examination. According to ICTV recommendations [45], all the TYLCSV sequences with gaps in homopolymeric regions were re-analyzed and gaps were discarded, but no additional variants were identified. Following the filtering steps, 10 full-length TYLCSV genomic variants were detected (Fig. 2A and Supplementary Table S3). All 10 variants contained a single SNP at the nucleotide (nt) position 2570 (2570-C/G) of TYLCSV C1 gene, which caused conservative serine to threonine (S/T) amino acid (aa) substitution in C1 protein (Fig. 2A). The most abundant virus variant (Variant-1) had only a single SNP (2570-C/G), while the other variants contained two to three SNPs, which were unequally distributed among C1, V2, C2, and C4 gene regions, preferentially affecting C1 and V2 genes. In summary, ten SNPs, affecting the C1, C2, and V2 genes, caused three synonymous and seven conservative amino acid substitutions in the corresponding proteins, while two SNPs, affecting C4 gene, caused nonconservative serine to leucine (S/L) and cysteine to stop codon substitutions (Fig. 2A). Interestingly, all the detected SNPs were unique and had no effect on the reconstructed TYLCSV consensus sequence, except for one SNP (2570-C/G), which was likely present in the original infectious clone sequence.

To validate the SNPs detected by the pipeline, the TYLCSV virus sequences amplified by RCA were re-analyzed using SRS technology. The existence of all SNPs was confirmed and their frequencies were calculated by mapping the reads to reference TYLCSV sequence (GenBank: Z25751.1). As expected, SNP 2570-C/G was present with a frequency of 100%, while other SNPs occurred with frequencies ranging from 1% to 7%, which highly correlated (correlation coefficient = 0.94) with the relative abundances of the corresponding virus variants detected by the pipeline (Fig. 2B). Thus, the same SNPs were detected with SRS at similar frequencies, while they cannot be assigned to individual haplotypes.

In addition to the full-length viral haplotypes, the pipeline enabled the profiling of partial/defective viral sequences. Four circular partial/defective viral sequences of varying lengths were identified after data processing (Supplementary Table S4). The most abundant defective sequence (Partial-1) covers 1948-nt region of the TYLCSV reference genome with 100% similarity and represents 0.6% of the most abundant variant (Variant-1) reads. Partial-1 sequence contains five open reading frames (ORFs) (Fig. 2C and Supplementary Fig. S2). The four ORFs encode the full-length V1, V2, C2, and C3 proteins, while the fifth ORF encodes the truncated version of C1 protein (131 aa) starting from the internal ATG codon. The V1 and V2 proteins are potentially expressed because the promoter region for rightward transcription remains intact. The leftward promoter region for transcription of truncated C1 protein and the replication initiation site are both absent which could make the molecule non-replicative. The second most abundant defective sequence (Partial-2) covers 1602-nt region of the TYLCSV reference genome with single SNP 2570-C/G and represents 0.5% of Variant-1 reads after processing. Partial-2 contains four ORFs encoding the full-length V1 and C4 proteins and truncated versions of V2 (131 aa) and C1 proteins (248 aa). The least abundant (0.2% frequency) defective molecules (Partial-3 and Partial-4) are 2631- and 1282-nt long and they encode five (V1, V2, C2, C3, and C4) and two (V1 and V2) full-length proteins, respectively. The Partial-3 defective molecule also contains an ORF that could express a truncated C1 protein (262 aa) (Fig. 2C and Supplementary Fig. S2). The production of Partial-1 and Partial-2 sequences was confirmed by short reads spanning the junction regions, whereas Partial-3 and Partial-4 sequences could not be confirmed by short reads, likely because of their low abundance (Supplementary Fig. S3).

To estimate the sensitivity of the designed pipeline with duplex reads selection, genomic variants and partial sequences were reconstructed from the Nanopore raw data followed by mapping to the reference sequence of TYLCSV virus clone, filtering the variants with deletions or/and insertions in homopolymeric regions and SNP analysis. After deconcatenation 4.6 times more TYLCSV viral reads were identified in the raw data compared with duplex data, that allowed to reconstruct 137 full-length and 25 partial TYLCSV virus sequences with no deletions or/and insertions in homopolymeric regions (Supplementary Tables S1 and S2), including ten genetic virus variants and four partial sequences detected in duplex data. While the pipeline enabled reconstruction of three additional



		Position	22	547	662	830	959	1362	1536	1897	1929	1962	2070	2392	2415	2439	2570	2752	
		AA Substitution	-	Arg/His	*Gly/Gly*	*Arg/Arg*	Gln/His	Ser/Asn	*Arg/Arg*	*Cys/Cys*	Ala/Thr	Glu/Gln	Phe/Leu	Ser/Leu	Asn/Tyr	Leu/Ile	Cys/Stop	Ser/Thr	-	
		SNP	C/T	G/A	G/A	G/A	G/T	C/T	C/T	G/A	C/T	C/G	A/G	G/A	T/A	G/T		C/G	T/A	
		Viral Gene / Variant	-	V2	V2	V2	V2	C3	C2	C1	C1	C1	C1	C4	C1	C1/C4		C1	-	SNPs
		TYLCSV-Z25751	C	G	G	G	G	C	C	G	C	C	A	G	T	G		C	T	-
Duplex data	Raw data	Variant-1	C	G	G	G	G	C	C	G	C	C	A	G	T	G		G	T	1
Duplex data	Raw data	Variant-2	C	G	G	G	G	C	C	G	C	G	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-3	C	A	G	G	G	C	C	G	C	C	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-4	T	G	G	G	G	C	C	G	C	C	A	A	T	G		G	T	3
Duplex data	Raw data	Variant-5	C	G	G	G	G	C	C	G	C	C	A	G	T	T		G	T	2
Duplex data	Raw data	Variant-6	C	G	G	A	G	C	C	G	C	C	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-7	C	G	G	G	G	C	C	G	T	C	A	G	A	G		G	T	3
Duplex data	Raw data	Variant-8	C	G	G	G	T	C	C	G	C	C	G	G	T	G		G	T	3
Duplex data	Raw data	Variant-9	C	G	G	G	G	C	C	A	C	C	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-10	C	G	G	G	G	C	T	G	C	C	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-11	C	G	G	G	G	C	C	G	C	C	A	G	T	G		G	A	2
Duplex data	Raw data	Variant-12	C	G	G	G	G	T	C	G	C	C	A	G	T	G		G	T	2
Duplex data	Raw data	Variant-13	C	G	A	G	G	C	C	G	C	C	A	G	T	G		G	T	2

Figure 2. (**A**) TYLCSV virus variants detected by the pipeline based on raw and duplex data. Viral genes and SNPs corresponding to these genes are color coded. Synonymous and non-synonymous (conservative and nonconservative) amino acid substitutions are depicted in italics and bold characters, respectively. (**B**) The relative frequencies of nine virus variants (Variant-2 to Variant-10) and corresponding SNPs detected in duplex data and three additional variants exclusively reconstructed from raw data (Variant-11 to Variant-13). The relative frequency of virus variants was calculated as a percentage of the most abundant TYLCSV virus variant (Variant-1; Supplementary Table S3). The relative frequencies of SNPs were obtained by mapping short reads to the TYLSCV Variant-1 sequence and mapping analysis by IGV. (**C**) TYLCSV partial genomic sequences. The diagram shows four TYLCSV partial genomic sequences with their sizes and locations based on the multiple sequence alignment with TYLCSV-Z25751 reference.

variants (Variant-11, Variant-12, and Variant-13) with SNPs confirmed by SRS (Fig. 2A and B) from raw data, the other 124 reconstructed virus variants containing 101 SNPs were not confirmed by SRS. Moreover, 11 out of 25 detected partial sequences contained unverified SNPs and were filtered out after processing (Supplementary Table S4). Thus, despite the higher sensitivity for virus variant reconstruction, the accuracy of the pipeline based on raw reads was not enough to distinguish between real mutations and sequencing errors without validation by SRS, while only "true" virus variants were reconstructed from high-quality duplex data.

In-depth profiling of multipartite DNA virus populations allows unambiguous identification of variants and genomic components generating partial viral genomes

The designed pipeline based on duplex data was applied for in-depth analysis of BBTV virus populations in three banana plant samples collected in DR Congo (BBTV-1), North Vietnam (BBTV-2), and South Vietnam (BBTV-3). BBTV duplex datasets were deconcatenated with TH using default parameters followed by sequence annotation and virus variant reconstruction with Genome Detective (Supplementary Table S5). After the selection of potential virus variant sequences without insertions in homopolymeric regions (23% for BBTV-1, 61% for BBTV-2, and 66% for BBTV-3), the full composition of BBTV components, including DNA-C, DNA-M, DNA-N, DNA-R, DNA-S, and DNA-U3, was reconstructed from all three datasets (Supplementary Table S6). BBTV sequences with deletions in homopolymeric regions were not detected. Subsequently, phylogenetic analysis of BBTV virus isolates was performed using the identified full-length genomic sequences of DNA-R components. The results confirmed the origins of BBTV-1, BBTV-2, and BBTV-3 virus isolates which were clustered with the previously identified BBTV isolates from DR Congo (GenBank: KM607637.1), North Vietnam (GenBank: AB113660.1), and South Vietnam (GenBank: AF416478.1), respectively (Supplementary Fig. S4).

Based on read counts of the identified BBTV virus components, DNA-C, DNA-S, and DNA-N appeared to be the most prevalent in BBTV-1 sample, while DNA-M/DNA-S/DNA-N and DNA-S/DNA-R/DNA-M were the top three viral components in BBTV-2 and BBTV-3 samples, respectively. Notably, the BBTV genomic component abundance was not consistent with the previously reported BBTV genome relative abundance [46], also known as the genomic formula [47,48]. Several factors could explain this discrepancy, including the impact of plant growth conditions or stage of infection on the genomic formula the presence of satellites in the samples that were analysed [46,49], as well as possible bias introduced by Phi-29 amplification and/or Nanopore sequencing [50].

The analysis revealed that DNA-C/DNA-N, DNA-M/DNA-N, and DNA-S/DNA-U3 were the most variable viral components in BBTV-1, BBTV-2, and BBTV-3 samples producing 12/7, 5/4, and 2/2 virus variants, respectively (Fig. 3A and Supplementary Table S6). The detected variants differ by one to three SNPs covering both translated and untranslated regions of viral components. Most of the SNPs caused synonymous or conservative aa substitutions, while four and one SNPs in BBTV-1 DNA-C (328-F/S; 388-A/V; 394–395-A/V) and DNA-N (706-C/F), respectively, and one SNP in BBTV-3 DNA-S (223-stop/Y) caused nonconservative aa substitutions (Supplementary Table S6). Notably, 223-stop/Y SNP caused ORF extension of DNA-S variant-2 in BBTV-3 sample. Similar extended ORFs of DNA-S have been identified in BBTV isolates from Africa (GenBank: GQ249344.1), India (GenBank: KT180280.1), Pakistan (GenBank: FJ859746.1), and Taiwan (GenBank: EF095164.1) (Supplementary Fig. S5). While all the detected SNPs in BBTV-2 and BBTV-3 samples were unique, three SNPs in BBTV-1 DNA-C and two SNPs in BBTV-1 DNA-N components were shared between nine and four variant sequences, respectively. Among them, two SNPs (394–395-A/V), which caused nonconservative aa substitutions in the translated region of BBTV-1 DNA-C component, were found in three variants, while the other three SNPs (946-A/C, 142-T/A, and 975-C/T) were identified in untranslated regions of six DNA-C and four DNA-N (two and two) variant sequences, respectively. In accordance with the results for monopartite TYLCSV, all the detected SNPs had no effect on the reconstructed BBTV consensus sequences. Overall, BBTV-1 sample showed more variability containing 6 major components and 21 full-genomic BBTV variants with relative abundance ranging from 1% to 48%, while BBTV-2 and BBTV-3 samples contained only six and two full-genomic variants with relative abundances ranging from 2% to 19% and 2% to 71%, respectively (Fig. 3A and Supplementary Table S6).

In addition to the BBTV viral components, the full-length genomic sequences of BBTV alphasatellites 2 and 3 (BBTA-2 and BBTA-3, family *Alphasatellitidae,* genus *Muscarsatellite*) and the recently discovered alphasatellite 5 (BBTA-5, family *Alphasatellitidae,* genus *Banaphisatellite*) were reconstructed from BBTV-2 and BBTV-3 datasets. While the sequences of BBTA-2 and BBTA-5 were detected in both datasets, BBTA-3 was exclusively reconstructed from BBTV-3. Notably, the detected sequences of BBTA-2 and BBTA-5 from North and South Vietnam were not identical, sharing only 93.5% and 96.1% homology, respectively (Supplementary Fig. S6).

The analysis of the potential BBTA variants revealed the presence of two variant sequences of BBTA-3 and BBTA-5 in the BBTV-3 sample, while other BBTA sequences did not show any variation. The BBTA-3 variant-2 had two SNPs in the protein coding region, which caused conservative valine to aspartic acid (V/D) and tyrosine to asparagine (Y/N) aa substitutions, while BBTA-5 variant-2 had a single SNP, which caused a synonymous (I/I) aa substitution (Supplementary Table S6).

As for the analysis of the monopartite virus, the pipeline allowed the reconstruction of partial genomic sequences of BBTV DNA virus components (Fig. 3B). Two partial sequences (DNA-N-partial-1 and DNA-S-partial-1), detected in the BBTV-2 sample, and one partial sequence (DNA-U3-partial-1) in BBTV-3 encoded intact ORFs with the potential to express full-length proteins, while five partial sequences (DNA-U3-partial-1 to DNA-U3-partial-3, DNA-N-partial-1 and DNA-N-partial-2) in BBTV-1 and four partial sequences (DNA-R-partial-1 to DNA-R-partial-4) in BBTV-2 and were truncated with potential short ORFs detected (Supplementary Tables S7 and S8).

In addition, partial sequences of BBTV alphasatellites could be profiled using the pipeline. Among them, five (BBTA-2-partial-1, BBTA-2-partial-3, BBTA-5-partial-1, BBTA-5-partial-2 and BBTA-5-partial-4) and seven (BBTA-2-partial-1 to BBTA-2-partial-4, and BBTA-3-partial-2 to BBTA-3-partial-4) partial sequences were found to encode intact ORFs in BBTV-2 and BBTV-3 samples, respectively, while the other four partial sequences (BBTA-2-partial-2, BBTA-2-partial-4, and BBTA-3-partial-1, BBTA-5-partial-3) were truncated with

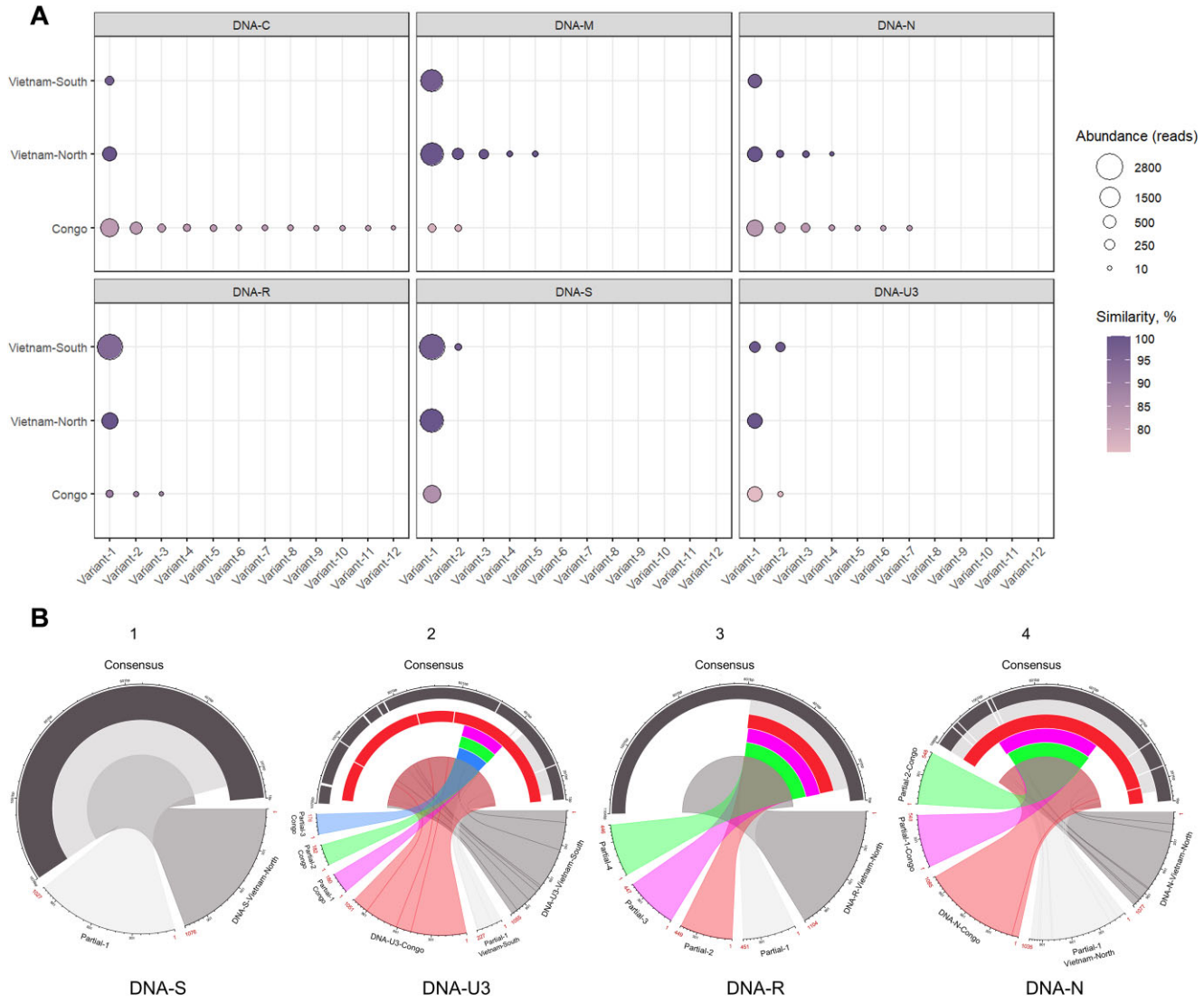


Figure 3. (**A**) BBTV virus variants, their read abundance, and similarity. The most abundant BBTV virus components detected in BBTV-2 (Vietnam-North) sample were used as references to calculate sequence identity (%). (**B**) Partial genomic sequences identified in BBTV-1 (Congo), BBTV-2 (Vietnam-North), and BBTV-3 (Vietnam-South) datasets and produced from (1) BBTV-Vietnam-North DNA-S, (2) BBTV-Congo and BBTV-Vietnam-South DNA-U3 (3) BBTV-Vietnam-North DNA-R, and (4) BBTV-Congo and BBTV-Vietnam-North DNA-N genomic components. SNPs are highlighted as white vertical lines in the consensus sequence.

potential short ORFs detected (Supplementary Tables S7 and S8). The partial sequences detected by the pipeline have not yet been reported based on a BLASTN search using the Nucleotide collection (nr/nt) database.

Discussion

In our work, we emphasized the usefulness of long-read sequencing technologies for in-depth analysis of plant virus populations to circumvent the limitations of SRS that allows determination of diversity, but does not permit the profiling of complete viral haplotypes. Our pipeline, based on circular DNA virus enrichment and high-quality sequencing data generation and analysis, allowed for the reconstruction of full-length genomic sequences of both monopartite and multipartite viral genomes. The sensitivity of the pipeline enabled detection of virus genomic variants with single nucleotide differences and >1% relative abundance.

In the present pipeline, Genome Detective online tool allows deconcatenation of RCA raw/duplex data, correction, and annotation of deconcatenated reads to output high-quality intact viral full-genome sequences and their relative abundances. The highest reliability of the analysis can be achieved by using duplex reads. The use of duplex reads increased the overall quality and accuracy of the analysis as illustrated by the detection of "true" TYLCSV virus variants which were confirmed by SGS. The pipeline generated numerous false positive virus variant sequences when using raw reads. Because the percentage of duplex reads can vary significantly between libraries, the use of total reads can be necessary to identify low abundance virus haplotypes for samples with low duplex reads but an independent validation approach would be required to validate the low abundance virus haplotypes.

Our analysis also showed the error-correction step performed by Genome Detective allows the accurate reconstruction of the original virus clone sequence without indels/deletions in homopolymeric regions. However, the difficulty of generating accurate sequencing of homopolymeric regions by TGS should not be underestimated and therefore

a careful examination of the viral sequence output remains necessary.

The sensitivity of the pipeline allowed the profiling of populations of mono- and multipartite DNA viruses, including the detection of single nucleotide variations in their full-length genome sequences with relative abundance as low as 1%. Using field samples of a multipartite virus, we could show that the pipeline was effective in detecting virus variants from the different genomic components. Noticeably, the pipeline allowed the unambiguous detection of partial virus genomic components which represents pioneering in-depth profiling of partial genomic components accumulating during infection. The output sequences of the partial genomic components indicate that they originate from the most abundant TYLCSV and BBTV virus variants. While the primary functions of partial sequences remain largely unknown, they are reported to affect the virus infection and is suggested to play an important role in virus genetic diversity [51–53]. Thus, the further analysis of the emerging virus populations in the context of different conditions, e.g. time, geographical origin, host plant, and environment, with our pipeline might shed light on the evolution and adaptation of the viruses to the host plant antiviral responses and/or changing environmental conditions. The unambiguous identification of virus variants in samples by our pipeline could also be instrumental to isolate variants causing mild symptoms for use in cross-protection approaches against the naturally severe parental viral isolates. The accurate profiling of circular DNA virus populations as performed by our pipeline, including low abundance variants and partial genomic components, holds great potential to monitor the emergence of virus variants in mono- and multipartite virus populations as well as to study the complex alteration of virus populations based on factors such as genetics of the host plant and changing environmental conditions.

Acknowledgements

We are grateful to Dr Jesús Navas-Castillo, Dr Elvira Fiallo Olivé, Dr Rony Swennen, and Dr Ha Viet Cuong for providing plant samples.

Author contributions: Victor Golyaev (Methodology, Formal analysis, Visualization, Writing—original draft, Writing—review & editing), Sam Dierickx (Methodology), Koen Deforche (Methodology), Wim Dumon (Supervision), and Hervé Vanderschuren (Conceptualization, Supervision, Project administration, Funding acquisition, Writing—review & editing)

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was funded by Horizon 2020 Framework Programme (Grant No. 101000570, VIRTIGATION) under the Horizon H2020-SFS-2020-2 program and the KU Leuven C1 funding scheme (Grant No. 3E210538). Funding to pay the Open Access publication charges for this article was provided by Horizon 2020 Framework Programme and the KU Leuven.

Data availability

Short read and Nanopore raw sequence data of TYLCSV and BBTV viruses are available at Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/PRJNA1113217).
Full length annotated genome sequences (TYLCSV-BBTV-Genome-Detective-variants.rar) and short read mapping files (TYLCSV-BGI-mapping.bam; TYLCSV-genome-Nanopore.fasta) are freely available at https://zenodo.org/records/11099227.

References

1. Adams I, Fox A. Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. In: Wang A, Zhou X (eds.), *Current Research Topics in Plant Virology*. Cham: Springer, 2016, 323–35
2. Rhie A, McCarthy SA, Fedrigo O *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Virus Evol* 2021;**592**:737–46. https://doi.org/10.1038/s41586-021-03451-0
3. Schmid M, Frei D, Patrignani A *et al.* Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res* 2018;**46**:8953–65. https://doi.org/10.1093/nar/gky726
4. Van Regenmortel MHV. Virus species and virus identification: past and current controversies. *Infect Genet Evol* 2007;**7**:133–44. https://doi.org/10.1016/j.meegid.2006.04.002
5. Maclot F, Candresse T, Filloux D *et al.* Illuminating an ecological blackbox: using high throughput sequencing to characterize the plant virome across scales. *Front Microbiol* 2020;**11**:578064. https://doi.org/10.3389/fmicb.2020.578064
6. Cai D, Sun Y. Reconstructing viral haplotypes using long reads. *Bioinformatics* 2022;**38**:2127–34.
7. Brait N, Külekçi B, Goerzer I. Long range PCR-based deep sequencing for haplotype determination in mixed HCMV infections. *BMC Genomics* 2022;**23**:31. https://doi.org/10.1186/s12864-021-08272-z
8. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. *PLoS One* 2021;**16**:e0257521. https://doi.org/10.1371/journal.pone.0257521
9. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. *Nat Rev Genet* 2023;**24**:627–41. https://doi.org/10.1038/s41576-023-00600-1
10. Hu T, Chitnis N, Monos D *et al.* Next-generation sequencing technologies: an overview. *Hum Immunol* 2021;**82**:801–11. https://doi.org/10.1016/j.humimm.2021.02.012
11. Zhang T, Li H, Ma S *et al.* The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. *Appl Environ Microb* 2023;**89**:e0060523. https://doi.org/10.1128/aem.00605-23
12. Yu SCY, Deng J, Qiao R *et al.* Comparison of single molecule, real-time sequencing and nanopore sequencing for analysis of the size, end-motif, and tissue-of-origin of long cell-free DNA in plasma. *Clin Chem* 2023;**69**:168–79. https://doi.org/10.1093/clinchem/hvac180
13. Rojas MR, Macedo MA, Maliano MR *et al.* World management of geminiviruses. *Annu Rev Phytopathol* 2018;**56**:637–77. https://doi.org/10.1146/annurev-phyto-080615-100327
14. Wang XW, Blanc S. Insect transmission of plant single-stranded DNA viruses. *Annu Rev Entomol* 2021;**66**:389–405. https://doi.org/10.1146/annurev-ento-060920-094531
15. Stenger DC, Revington GN, Stevenson MC *et al.* Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc*

Natl Acad Sci USA 1991;**88**:8029–33. https://doi.org/10.1073/pnas.88.18.8029
16. Bonnamy M, Blanc S, Michalakis Y. Replication mechanisms of circular ssDNA plant viruses and their potential implication in viral gene expression regulation. *mBio* 2023;**14**:e01692-23. https://doi.org/10.1128/mbio.01692-23
17. Jenkins GM, Rambaut A, Pybus OG *et al.* Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* 2002;**54**:156–65. https://doi.org/10.1007/s00239-001-0064-3
18. Wu M, Wei H, Tan H *et al.* Plant DNA polymerases α and δ mediate replication of geminiviruses. *Nat Commun* 2021;**12**:2780. https://doi.org/10.1038/s41467-021-23013-2
19. Ortega-Del Campo S, Grigoras I, Timchenko T *et al.* Twenty years of evolution and diversification of digitaria streak virus in *Digitaria setigera*. *Virus Evol* 2021;**7**:veab083. https://doi.org/10.1093/ve/veab083
20. LaTourrette K, Garcia-Ruiz H. Determinants of virus variation, evolution, and host adaptation. *Pathogens* 2022;**11**:1039.
21. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev* 2012;**76**:159–216. https://doi.org/10.1128/MMBR.05023-11
22. Grande-Pérez A, Gómez-Mariano G, Lowenstein PR *et al.* Mutagenesis-induced, large fitness variations with an invariant arenavirus consensus genomic nucleotide sequence. *J Virol* 2005;**79**:10451–9. https://doi.org/10.1128/JVI.79.16.10451-10459.2005
23. Sánchez-Campos S, Domínguez-Huerta G, Díaz-Martínez L *et al.* Differential shape of geminivirus mutant spectra across cultivated and wild hosts with invariant viral consensus sequences. *Front Plant Sci* 2018;**9**:932. https://doi.org/10.3389/fpls.2018.00932
24. Kolmogorov M, Bickhart DM, Behsaz B *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10. https://doi.org/10.1038/s41592-020-00971-x
25. Koren S, Walenz BP, Berlin K *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**:722–36. https://doi.org/10.1101/gr.215087.116
26. Chen Y, Nie F, Xie SQ *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* 2021;**12**:60. https://doi.org/10.1038/s41467-020-20236-7
27. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 2016;**32**:2103–10.
28. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**:155–8. https://doi.org/10.1038/s41592-019-0669-3
29. Knyazev S, Tsyvina V, Shankar A *et al.* Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Res* 2021;**49**:e102. https://doi.org/10.1093/nar/gkab576
30. Posada-Céspedes S, Seifert D, Topolsky I *et al.* V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* 2021;**37**:1673–80.
31. Eliseev A, Gibson KM, Avdeyev P *et al.* Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect Genet Evol* 2020;**82**:104277. https://doi.org/10.1016/j.meegid.2020.104277
32. Luo X, Kang X, Schönhuth A. Strainline: full-length *de novo* viral haplotype reconstruction from noisy long reads. *Genome Biol* 2022;**23**:29. https://doi.org/10.1186/s13059-021-02587-6
33. Cai D, Shang J, Sun Y. HaploDMF: viral haplotype reconstruction from long reads via deep matrix factorization. *Bioinformatics* 2022;**38**:5360–7.
34. Gao Y, Liu B, Wang Y *et al.* TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* 2019;**35**:i200–7.
35. Vilsker M, Moosa Y, Nooij S *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* 2019;**35**:871–3.
36. Mehta D, Cornet L, Hirsch-Hoffmann M *et al.* Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq. *Nat Protoc* 2020;**15**:1673. https://doi.org/10.1038/s41596-020-0301-0
37. Chen Y, Chen Y, Shi C *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* 2018;**7**:gix120. https://doi.org/10.1093/gigascience/gix120
38. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.
39. Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6. https://doi.org/10.1038/nbt.1754
40. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257. https://doi.org/10.1186/s13059-019-1891-0
41. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–8. https://doi.org/10.1038/s41592-021-01101-x
42. Deforche K. An alignment method for nucleic acid sequences against annotated genomes. bioRxiv, https://doi.org/10.1101/200394, 11 October 2017, preprint: not peer reviewed.
43. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
44. Bonfield JK, Marshall J, Danecek P *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience* 2021;**10**:giab007. https://doi.org/10.1093/gigascience/giab007
45. Brown JK, Zerbini FM, Navas-Castillo J *et al.* Revision of begomovirus taxonomy based on pairwise sequence comparisons. *Arch Virol* 2015;**160**:1593–619. https://doi.org/10.1007/s00705-015-2398-y
46. Guyot V, Rajeswaran R, Chu HC *et al.* A newly emerging alphasatellite affects banana bunchy top virus replication, transcription, siRNA production and transmission by aphids. *PLoS Pathog* 2022;**18**:e1010448. https://doi.org/10.1371/journal.ppat.1010448
47. Sicard A, Yvon M, Timchenko T *et al.* Gene copy number is differentially regulated in a multipartite virus. *Nat Commun* 2013;**4**:2248. https://doi.org/10.1038/ncomms3248
48. Bonnamy M, Brousse A, Pirolles E *et al.* The genome formula of a multipartite virus is regulated both at the individual segment and the segment group levels. *PLoS Pathog* 2024;**20**:e1011973. https://doi.org/10.1371/journal.ppat.1011973
49. Mansourpour M, Gallet R, Abbasi A *et al.* Effects of an alphasatellite on the life cycle of the nanovirus *Fababeannecroticyellowsvirus*. *J Virol* 2022;**96**:e0138821. https://doi.org/10.1128/JVI.01388-21
50. Boezen D, Johnson ML, Grum-Grzhimaylo AA *et al.* Evaluation of sequencing and PCR-based methods for the quantification of the viral genome formula. *Virus Res* 2023;**326**:199064. https://doi.org/10.1016/j.virusres.2023.199064
51. Ndunguru J, Legg JP, Fofana IBF *et al.* Identification of a defective molecule derived from DNA-A of the bipartite begomovirus of East African cassava mosaic virus. *Plant Pathol* 2006;**55**:2–10. https://doi.org/10.1111/j.1365-3059.2005.01289.x
52. Patil BL, Dasgupta I. Defective interfering DNAs of plant viruses. *Crit Rev Plant Sci* 2006;**25**:47–64. https://doi.org/10.1080/07352680500391295
53. Stainton D, Martin DP, Muhire BM *et al.* The global distribution of *banana bunchy top virus* reveals little evidence for frequent recent, human-mediated long distance dispersal events. *Virus Evol* 2015;**1**:vev009. https://doi.org/10.1093/ve/vev009

Received: August 3, 2024. **Revised:** February 19, 2025. **Editorial Decision:** March 9, 2025. **Accepted:** March 11, 2025