

RESEARCH

Open Access



Assessing the de novo assemblers: a metaviromic study of apple and first report of citrus concave gum-associated virus, apple rubbery wood virus 1 and 2 infecting apple in India

Zainul A. Khan^{1,4†}, Susheel Kumar Sharma^{1*†}, Nitika Gupta^{1†}, Damini Diksha^{1†}, Pooja Thapa^{1†}, Mailem Yazing Shimray¹, Malyaj R. Prajapati¹, Sajad U. Nabi², Santosh Watpade³, Mahendra K. Verma² and Virendra K. Baranwal^{1*}

Abstract

Background The choice of de novo assembler for high-throughput sequencing (HTS) data remains a pivotal factor in the HTS-based discovery of viral pathogens. This study assessed de novo assemblers, namely Trinity, SPAdes, and MEGAHIT for HTS datasets generated on the Illumina platform from 23 apple samples, representing 15 exotic and indigenous apple varieties and a rootstock. The assemblers were compared based on assembly quality metrics, including the largest contig, total assembly length, genome coverage, and N50.

Results MEGAHIT was most efficient assembler according to the metrics evaluated in this study. By using multiple assemblers, near-complete genome sequences of citrus concave gum-associated virus (CCGaV), apple rubbery wood virus 1 (ARWV-1), ARWV-2, apple necrotic mosaic virus (ApNMV), apple mosaic virus, apple stem pitting virus, apple stem grooving virus, apple chlorotic leaf spot virus, apple hammerhead viroid and apple scar skin viroid were reconstructed. These viruses were further confirmed through Sanger sequencing in different apple cultivars. Among them, CCGaV, ARWV-1 and ARWV-2 were recorded from apples in India for the first time. The analysis of virus richness revealed that ApNMV was dominant, followed by ARWV-1 and CCGaV. Moreover, MEGAHIT identified novel single-nucleotide variants.

Conclusions Our analyses highlight the crucial role of assembly methods in reconstructing near-complete apple virus genomes from the Illumina reads. This study emphasizes the significance of employing multiple assemblers for de novo virus genome assembly in vegetatively propagated perennial fruit crops.

[†]Zainul A. Khan, Susheel Kumar Sharma, Nitika Gupta, Damini Diksha and Pooja Thapa contributed equally to this work.

*Correspondence:

Susheel Kumar Sharma
susheelsharma19@gmail.com
Virendra K. Baranwal
vbaranwal2001@yahoo.com

Full list of author information is available at the end of the article



Keywords Apple virome, High-throughput sequencing, De novo assemblers, Citrus concave gum-associated virus, Apple rubbery wood virus 1, Single-nucleotide variants

Background

Apple (*Malus domestica*) is an economically important fruit crop grown in temperate regions of the world, including India. India is the 5th largest apple producing country in the world and produced 2,589,000 metric tons of apples from a 315,000 ha area in 2021–2022 (<https://agriwelfare.gov.in/en/StatHortEst>, Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India). Jammu and Kashmir and Himachal Pradesh are the major apple growing states of India.

Globally, over twenty viruses and viroids infecting apple trees have been documented [1, 2]. Apple mosaic virus (ApMV), apple necrotic mosaic virus (ApNMV), apple chlorotic leaf spot virus (ACLSV), apple stem grooving virus (ASGV), apple stem pitting virus (ASPV), apple rubbery wood virus 1 (ARWV-1), ARWV-2, citrus concave gum-associated virus (CCGaV), apple scar skin viroid (ASSVd) and apple hammerhead viroid (AHVd) are of possible economic importance [1, 2]. Apple trees are vegetatively propagated that allows the transmission and accumulation of viruses through cuttings and grafting practices. The latent viruses such as ASGV, ASPV and ACLSV don't produce symptoms in most commercially grown apple cultivars [3, 4]. In susceptible varieties, however, these viruses produce various symptoms, such as stem grooving, malformation of young leaves, leaf epinasty or deformation at top grafting unions. Leaf mosaic and dappled fruit symptoms are caused by ApNMV and ASSVd in apple cultivars, respectively [5, 6].

During infection, RNA viruses naturally accumulate random genetic variants due to their mutation rates being up to a million times higher than those of their hosts [7]. This high mutation rate increases their pathogenicity and ability to evolve into new species, leading to quasispecies diversity [8, 9]. The viral quasispecies, a collection of various variants, determine the properties of the viral population [10]. High-throughput sequencing (HTS) can identify unknown or novel viruses using de novo assemblers and study the virus variants. Virome studies rely on de novo assembly methods because of uncharacterized viral sequences and the absence of a universal marker gene. Metagenomic assemblers use de Bruijn graph (DBG) approaches to efficiently handle complex datasets, however, microbial metagenomes present challenges that can complicate DBG assembly processes, leading to fragmentation and misassembly [11, 12]. Virome data includes viral genomes with abundant repeat regions,

hypervariable regions associated with host interactions, and high mutation rates, causing increased metagenomic complexity and strain diversity [13].

The extent of variations in assembly quality among different assemblers has not been previously documented for plant viruses infecting vegetatively propagated perennial fruit crops. The present study provides a thorough analysis of the effectiveness of three de novo assemblers for virus genome assembly using 23 paired-end Illumina datasets generated from apple trees and nursery plants. We employed various assembly quality metrics, such as the largest contig, overall assembly length, genome coverage percentage and N50, to compare assemblers. It was found that the number of variants in contigs retrieved using different assemblers varies significantly, and that there were differences in variants between assemblers for the same sample. For the first time, we report the identification and detection of CCGaV, ARWV-1, and ARWV-2 from apple trees in India. Additionally, we identified several other viruses and viroids infecting apple plants, including ASPV, ASGV, ACLSV, ApNMV, ApMV, ASSVd, and AHVd.

Results

Sample preparation, library construction and RNA sequencing


To study virus infection in apple trees, twenty-three samples from 15 cultivars were examined. Ten apple samples showed mosaic symptoms, while the remaining samples were asymptomatic (Table 1). The statistical HTS datasets, including raw sequence reads, high-quality reads, host unaligned reads, Q20, Q30 and GC percentage are detailed in Table 2.

Performance of different de novo assemblers

Assembly quality

To assess assembly quality across various assemblers, two de novo transcriptome assemblers (Trinity and SPAdes) and a de novo metagenome assembler (MEGAHIT) were selected (Table S1). The BLASTn results showed the identification of near-complete genome of ACLSV, ASGV, ASPV, ApNMV, CCGaV, ApMV, ARWV-1, ARWV-2, AHVd and ASSVd using all three assemblers. MEGAHIT produced the largest number of contigs (183,901), whereas SPAdes generated the fewest (10,332), indicating a substantial 17-fold difference (Table S1). The recovery of genome fractions varied greatly among

Table 1 List of symptomatic and asymptomatic apple samples of different cultivars with their HTS code collected from Himachal Pradesh and Jammu and Kashmir

Location	HTS code and apple cultivar	Symptoms	Leaf sample	Location	HTS code and apple cultivar	Symptoms	Leaf sample
ICAR-IARI, Shimla, Himachal Pradesh, India	A1, Golden Delicious	Mosaic		ICAR-CITH, Srinagar, Jammu and Kashmir, India	A13, Starkrimson	Asymptomatic	
	A2, Red Fuji	Mosaic			A14, Oregon Spur	Asymptomatic	
	A3, Royal Delicious	Mosaic			A15, Oregon Spur	Mosaic	
ICAR-CITH, Srinagar, Jammu and Kashmir, India	A4, Red Gravenstein	Mosaic	NA ^a	SKUAST, Srinagar, Jammu and Kashmir, India	A16, M9 rootstock	Asymptomatic	
	A5, Gala Mast	Mosaic			A17, Golden Delicious	Mosaic	
Nursery, Kulgam, Jammu and Kashmir, India	A6, Gala Mast	Asymptomatic		SKUAST, Srinagar, Jammu and Kashmir, India	A18, Golden Delicious	Asymptomatic	
	A7, Spartan	Mosaic			A19, American Apirouge	Mosaic	
	A8, Spartan	Asymptomatic			Nursery, Kulgam, Jammu and Kashmir, India	A20, Coe Red Fuji	Asymptomatic
	A9, Gala Red Lum	Asymptomatic		A21, Gala SchniCo Red		Asymptomatic	
	A10, Ambri	Asymptomatic		A22, King Roat		Asymptomatic	
	A11, Ambri	Mosaic	NA ^a	Nursery, Kulgam, Jammu and Kashmir, India	A23, Z1	Asymptomatic	
	A12, Coe Red Fuji	Asymptomatic					

^a NA not available

Table 2 Summary of read statistics and host unaligned reads of 23 apple high throughput sequencing datasets

Apple samples	Statistics term													
	BioProject	BioSample	SRA accession	Total raw reads	Read length	Total filtered reads (HQ)	Filtered read length (HQ)	HQ reads (%)	Host aligned reads	Host unaligned reads	Aligned reads (%)	Q20 (%)	Q30 (%)	GC content (%)
A1	PRJNA1025925	SAMN37730397	SRR26446885	79,450,612	150	77,252,536	149.53	97.23	55,218,776	22,033,760	71.48	97.45	92.76	44.15
A2		SAMN37730398	SRR26446884	47,044,624	150	45,515,404	149.39	96.74	29,700,148	15,815,256	65.25	97.60	93.16	45.09
A3		SAMN37730399	SRR26446883	50,962,016	150	48,885,412	149.54	95.92	31,352,162	17,533,250	64.13	96.80	91.49	43.50
A4	PRJNA1032713	SAMN38003657	SRR26560097	79,960,516	150	76,974,898	149.82	96.26	53,236,292	23,738,606	69.16	98.33	94.57	44.61
A5		SAMN38003658	SRR26560096	51,576,664	150	50,176,138	149.68	97.28	33,727,213	16,448,925	67.22	97.37	92.65	43.01
A6		SAMN38003659	SRR26560095	72,347,404	150	69,191,646	149.44	95.63	45,262,457	23,929,189	65.42	97.54	93.35	46.56
A7		SAMN38003660	SRR26560094	50,314,234	150	48,509,994	149.82	96.41	33,862,433	14,647,561	69.81	97.34	92.51	42.97
A8		SAMN38003661	SRR26560093	59,158,860	150	56,671,812	149.48	95.79	37,537,930	19,133,882	66.24	97.29	92.44	42.47
A9		SAMN38003662	SRR26560922	89,139,454	150	85,558,178	149.65	95.98	56,986,461	28,571,717	66.61	97.78	93.77	42.77
A10		SAMN38003663	SRR26560921	44,232,822	150	42,532,292	149.61	96.15	29,141,368	13,390,924	68.52	97.96	93.51	42.66
A11		SAMN38003664	SRR26560920	89,531,714	150	85,427,860	149.82	95.41	59,628,591	25,799,269	69.80	98.17	94.19	44.13
A12		SAMN38003665	SRR26560919	48,572,308	150	47,538,670	149.19	97.87	33,313,869	14,224,801	70.08	98.14	93.96	42.80
A13		SAMN38003666	SRR26560918	59,674,798	150	57,801,336	149.74	96.86	40,242,506	17,558,830	69.62	98.32	94.60	43.03
A14		SAMN38003667	SRR26592227	55,488,456	150	53,690,774	149.76	96.76	37,543,002	16,147,772	69.92	98.10	93.88	42.94
A15		SAMN38003668	SRR26592226	60,418,336	150	58,494,966	149.44	96.81	40,403,583	18,091,383	69.07	98.00	93.57	42.51
A16		SAMN38003669	SRR26592225	55,416,230	150	54,046,420	149.82	97.52	36,198,840	17,847,580	66.98	97.49	92.84	42.49
A17		SAMN38003670	SRR26592224	44,263,552	150	43,248,384	149.32	97.70	29,963,466	13,284,918	69.28	98.08	93.78	42.84
A18	PRJNA1034804	SAMN38074578	SRR266638234	78,154,418	150	75,871,906	149.56	97.07	68,106,172	7,765,734	89.76	97.90	93.85	54.09
A19		SAMN38074579	SRR266638233	42,260,246	150	41,283,670	149.38	97.68	26,527,398	14,756,272	64.26	98.11	93.87	42.42
A20	PRJNA1034824	SAMN38075211	SRR26661760	43,355,420	150	41,724,262	149.47	96.23	26,491,768	15,232,494	63.49	98.01	93.61	41.97
A21		SAMN38075212	SRR26661759	72,960,628	150	68,774,874	149.56	94.26	49,014,767	19,760,107	71.27	98.07	93.94	46.82
A22		SAMN38075213	SRR26661758	94,144,834	150	90,124,434	149.36	95.72	66,561,608	23,562,826	73.86	97.64	92.83	47.51
A23		SAMN38075214	SRR26661757	69,987,994	150	66,964,376	149.83	95.67	45,126,024	21,838,352	67.39	98.11	93.88	42.82

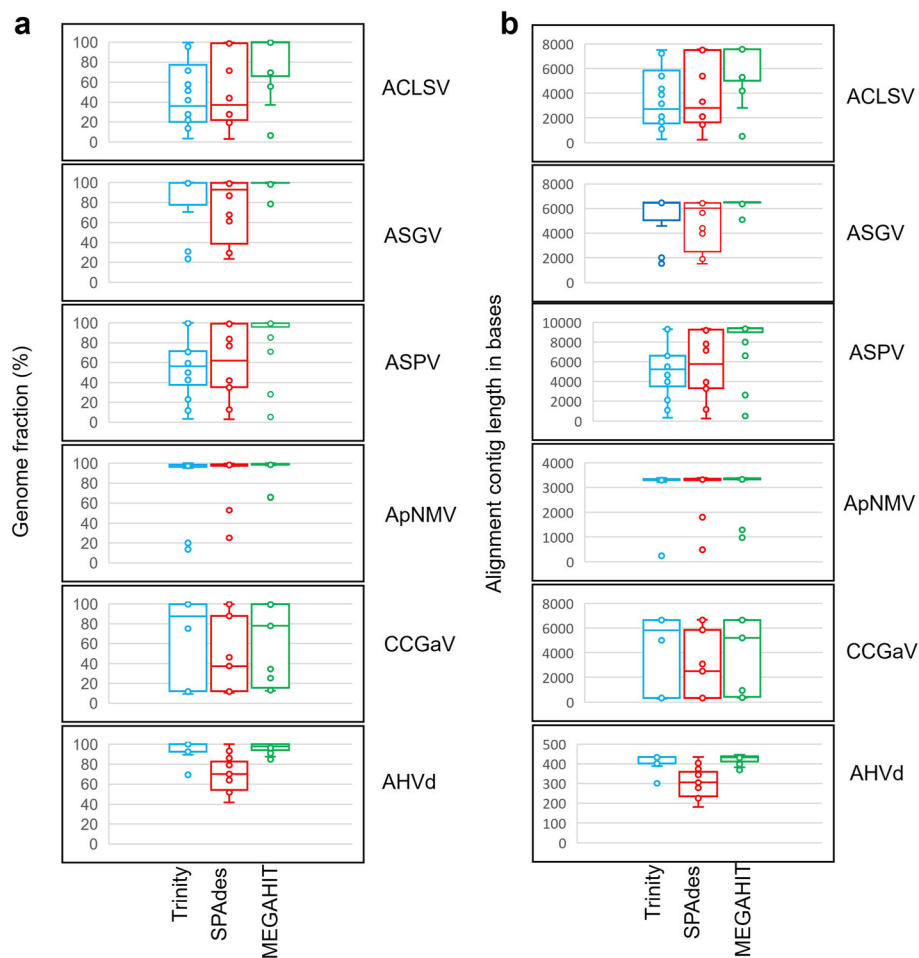


Fig. 1 Comparison of de novo assemblers (Trinity, SPAdes and MEGAHIT) for the reconstruction of virus and viroid genome sequences. **A** Fraction of virus and viroid genome assembly by three assemblers. **B** Alignment of longest virus and viroid contigs assembled by different assemblers. ACLSV, apple chlorotic leaf spot virus; ASGV, apple stem grooving virus; ASPV, apple stem pitting virus; ApNMV, apple necrotic mosaic virus; CCGaV, citrus concave gum-associated virus; AHVd, apple hammerhead viroid

assemblers (Fig. 1a and Fig. S1). MEGAHIT recovered a large fraction (>99%) of ACLSV, ASGV, ASPV, ApNMV, and CCGaV genomes in 10, 10, 14, 6, and 5 libraries, respectively (Fig. 1a and Table S2). Conversely, SPAdes assembled a substantial genome fraction (>99%) in 3, 5, 5, 3, and 2 libraries for ACLSV, ASGV, ASPV, ApNMV, and CCGaV, respectively (Fig. 1a and Table S2). On the other hand, Trinity achieved a larger genome fraction (>99%) of ACLSV, ASGV, ASPV, ApNMV, and CCGaV genomes in 2, 9, 2, 1, and 5 libraries, respectively (Fig. 1a and Table S2). Trinity, MEGAHIT, and SPAdes successfully reconstructed a substantial portion (100%) of the AHVd genome in 11, 7, and 1 library, respectively, among the 15 HTS libraries where AHVd contigs were identified (Fig. 1a and Table S2). ARWV-1 and ASSVd contigs appeared in two HTS datasets, whereas ApMV and ARWV-2 contigs were found in just one. ApMV showed a genome fraction (>99%) from SPAdes and MEGAHIT

assembly, while MEGAHIT achieved a large fraction recovery (>99%) of ARWV-2. For ARWV-1, Trinity and SPAdes assembled a large fraction (>99%), and Trinity alongside MEGAHIT assembled large fraction (100%) of ASSVd genome (Fig. S1a and Table S2).

MEGAHIT achieved genome recovery rates of 71% (10/14) for ACLSV, 83% (10/12) for ASGV, 77.8% (14/18) for ASPV, 50% (6/12) for ApNMV, and 55.6% (5/9) for CCGaV, all with genome coverage exceeding 99% by a single contig. Similarly, Trinity demonstrated genome recovery rates of 14.3% (2/14) for ACLSV, 75% (9/12) for ASGV, 11.1% (2/18) for ASPV, 8.3% (1/12) for ApNMV, and 55.6% (5/9) for CCGaV, with genome coverage surpassing 99% through a single contig. SPAdes, on the other hand, exhibited genome recovery rates of 21.4% (3/14) for ACLSV, 41.7% (5/12) for ASGV, 27.8% (5/18) for ASPV, 25% (3/12) for ApNMV, and 22.2% (2/9) for CCGaV, all

through a single contig while achieving genome coverage exceeding 99% (Table S2).

MEGAHIT was more efficient in assembling ACLSV, ASGV, ASPV, ApNMV, ApMV and ARWV-2 contigs, while all three assemblers were complementary in assembling CCGaV, ARWV-1, AHVd and ASSVd contigs (Fig. 1b, Fig. S1b and Table S3).

Quality matrices

The N50 values ranged from 305 to 1,118 (Fig. S2 and Table S1). The smallest N50 value (305) was identified in the A21 HTS dataset assembled by Trinity, while the largest N50 value (1,118) was identified in the A4 HTS dataset assembled by MEGAHIT. Specifically, Trinity N50 values ranged from 305 in the A21 HTS dataset to 553 in the A23 HTS dataset. SPAdes N50 values ranged from 389 in the A6 HTS dataset to 840 in the A2 HTS dataset. MEGAHIT provided better N50 values which ranged from 849 in the A21 HTS dataset to 1,118 in the A4 HTS dataset (Fig. S2 and Table S1).

Virus/viroid genome coverage

The virus and viroid contigs assembled by Trinity, SPAdes and MEGAHIT were compared in 23 HTS datasets. While MEGAHIT successfully achieved greater genome coverage and longer contig lengths of ASGV, ACLSV, ASPV, ApNMV, CCGaV, ApMV, and AHVd, it encountered challenges in assembling ASSVd, CCGaV RNA1, and ARWV-1 segment S in a single apple HTS sample (Table S4).

Perceived viral richness

The impact of different assemblers on virus copy numbers and fragments per kilobase of transcript per million mapped reads (FPKM) values was assessed using three sets of HTS data (Fig. 2a and b). The virus contigs identified from Trinity, SPAdes, or MEGAHIT were employed as reference genomes for read mapping. The mapped reads were then utilized to calculate virus abundance within the HTS datasets. In most instances, the choice of assemblers had no significant effect on virus copy numbers. However, in few cases, both SPAdes and MEGAHIT

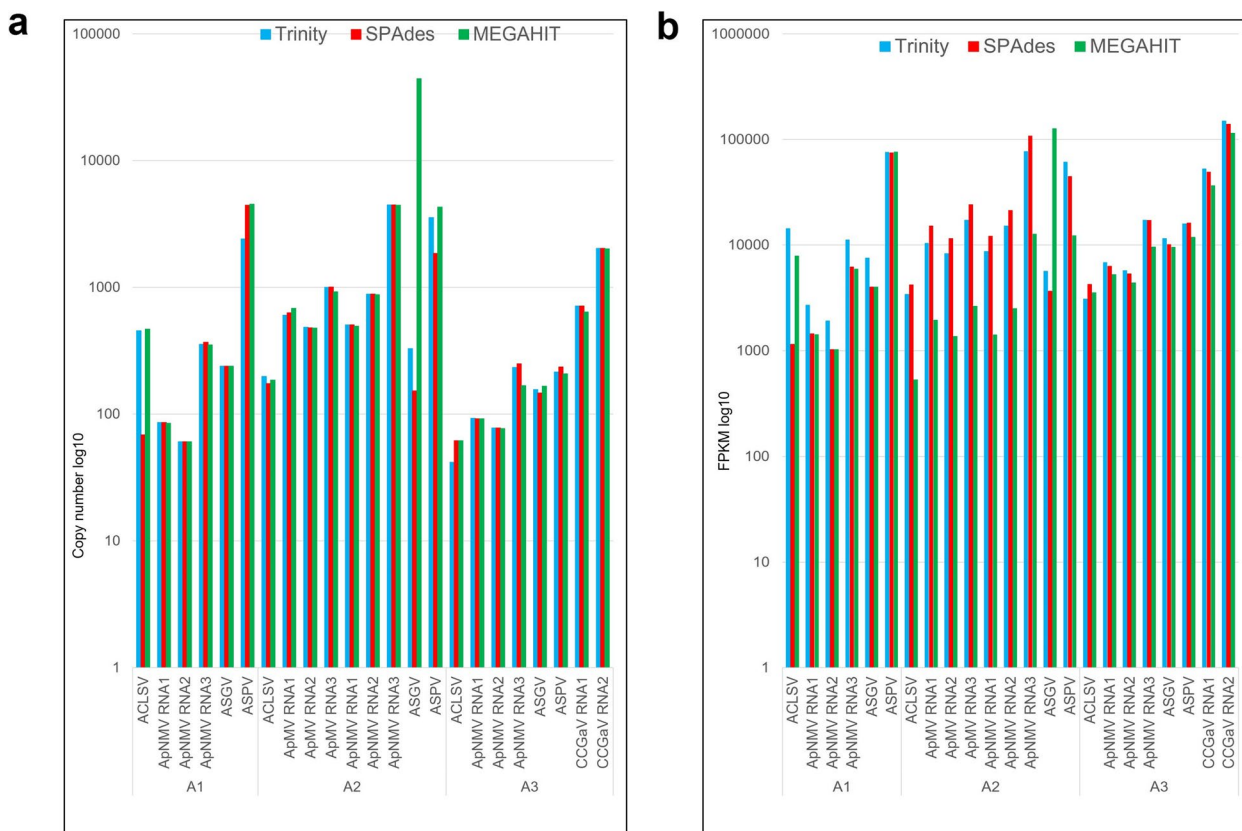


Fig. 2 Comparison of virus copy numbers (A) and FPKM (B) in three apple HTS samples. The virus contigs assembled through three different assemblers including Trinity, SPAdes and MEGAHIT were used as reference to calculate individual virus copies and FPKM in each apple sample. ACLSV, apple chlorotic leafspot virus; ASGV, apple stem grooving virus; ASPV, apple stem pitting virus; ApNMV, apple necrotic mosaic virus; ApMV, apple mosaic virus; CCGaV, citrus concave gum-associated virus

assembled contigs exhibited higher viral copy numbers compared to those assembled by Trinity. Notably, the MEGAHIT assembly of the ASGV contig in the A2 HTS dataset displayed the highest copy numbers (Fig. 2a). In contrast, the Trinity-assembled CCGaV RNA1 in the A3 HTS dataset showed the highest FPKM values (Fig. 2b). Furthermore, among the 23 apple HTS datasets, ApNMV RNA1 exhibited the highest copies (15,446.55) and FPKM values (345,656.26) in A20 and A19, respectively for viruses, while AHVd displayed the highest copies (78,992.41) and FPKM values (2,301,912.81) in A5 and A12, respectively for viroids (Table S5).

Single nucleotide variants (SNVs)

To explore unique variants specific to each assembler, a comparison of the variants identified by the three assemblers was conducted. We analysed the impact of assemblers on virus SNVs in five apple HTS datasets. ApNMV RNA1 contigs assembled through Trinity, SPAdes and MEGAHIT were selected. In two sets of HTS data, a consistent number of variants across all three assemblers was observed. However, unique SNVs were identified in two HTS samples when SPAdes and MEGAHIT were utilized. In one of the HTS datasets, unique SNVs were

specifically observed when the MEGAHIT assembler was used (Fig. 3a). Trinity, SPAdes and MEGAHIT identified 207 and 28 SNVs in ApNMV RNA1 in the A2 and A3 HTS datasets, respectively (Fig. 3a). In the A1 HTS dataset, two distinct SNVs (G>C and G>C) were detected exclusively when employing the MEGAHIT assembler. In the A4 HTS dataset, the use of both SPAdes and MEGAHIT revealed three unique SNVs (T>A, T>A, and G>A) (Fig. 3b). Additionally, in the A5 HTS dataset, SPAdes and MEGAHIT identified an additional 16 and 18 SNVs, respectively (Fig. 3a and b). Moreover, among the 23 apple HTS datasets, ASPV exhibited highest number of SNVs (813) in A1 (Table S5).

Phylogenetic analysis

Phylogenetic analysis showed that the CCGaV identified in A3, A5, A9, and A21 HTS data shared a close genetic relationship with CCGaV previously reported in apples from Brazil. In contrast, the CCGaV identified in A17 and A23 HTS data exhibited a close relationship with the CCGaV reported in apples from Italy (Fig. 4a). ARWV-1 identified in the present study in A13 HTS data grouped together with Belgium isolate of ARWV-1 (OK398019) (Fig. 4b). Further, the ApNMV found in 11 HTS datasets

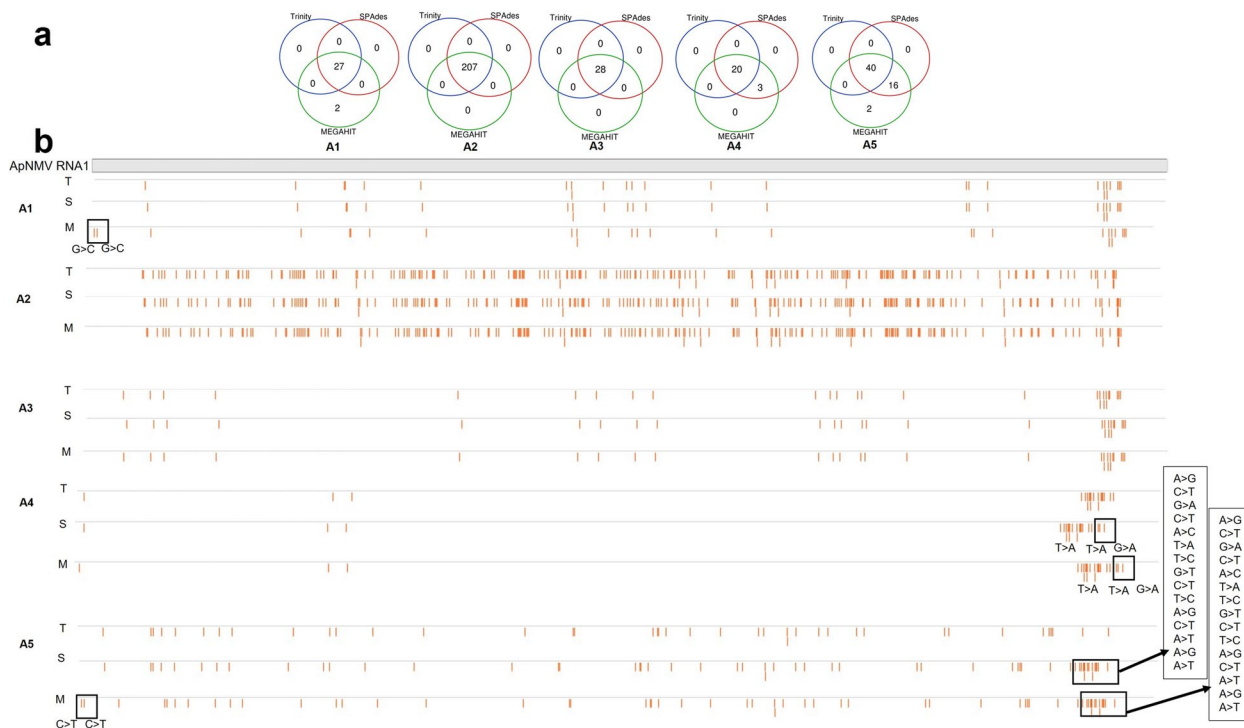


Fig. 3 Genomic variants discrepancies among three de novo assemblers (Trinity, SPAdes and MEGAHIT). **A** Unique and overlap of variants of apple necrotic mosaic virus (ApNMV) RNA1 among Trinity, SPAdes and MEGAHIT assemblies in five apple HTS samples including single nucleotide variants (SNVs). **B** Identification of common and novel SNVs in ApNMV RNA1 from five apple HTS datasets. Unique SNVs are highlighted in boxes along with their corresponding nucleotide substitutions. Grey filled box shows ApNMV RNA1 genome and brown bars represent SNVs. T, Trinity; S, Spades; M, MEGAHIT

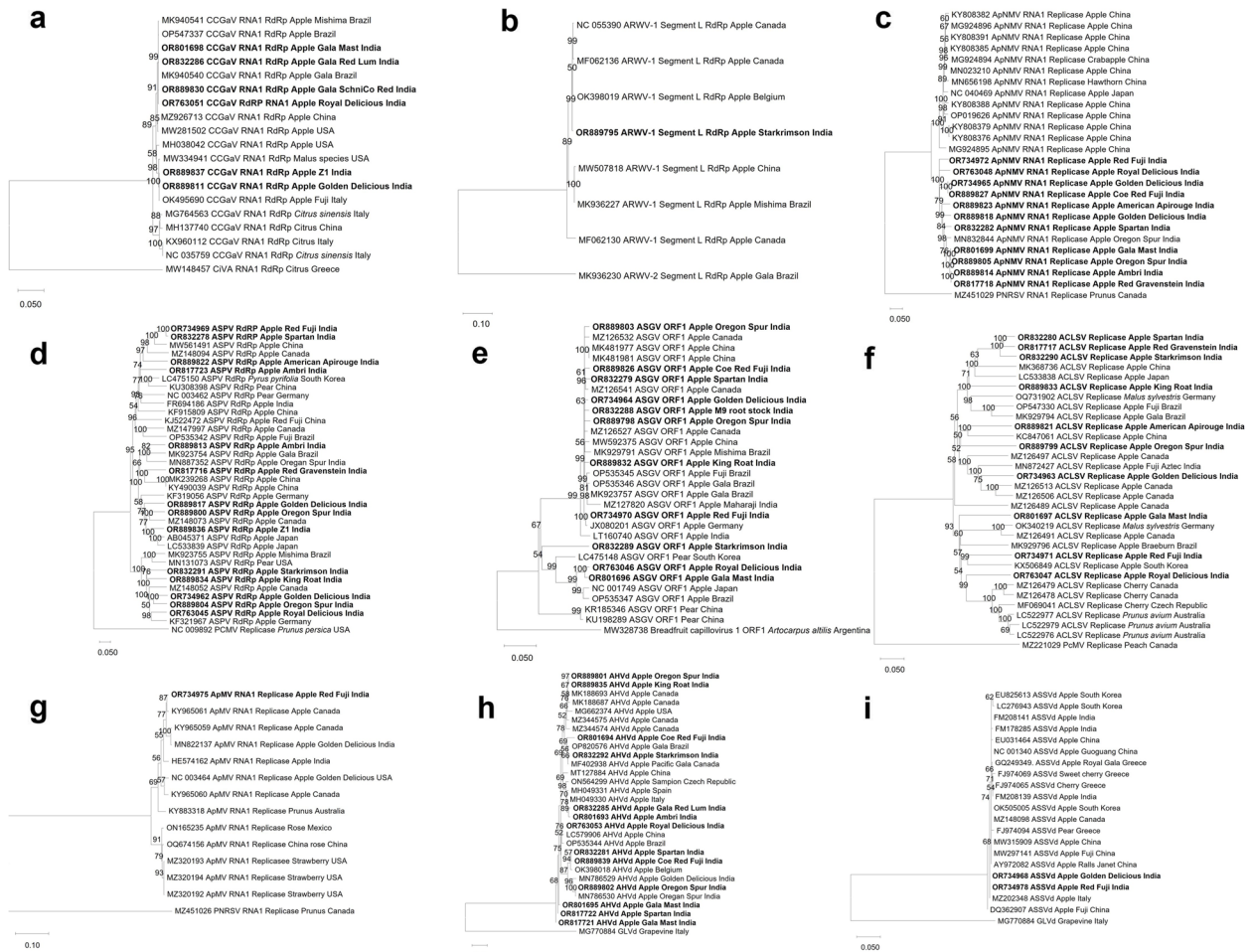


Fig. 4 Phylogenetic trees showing the relationship of **(A)** RdRp of CCGaV, **(B)** RdRp of ARWV-1, **(C)** replicase of ApNMV, **(D)** RdRp of ASPV, **(E)** RdRp of ASGV, **(F)** replicase of ACLSV, **(G)** replicase of ApMV, **(H)** AHVd and **(I)** ASSVd, under study (bold fonts) with other isolates of CCGaV, ARWV-1, ApNMV, ASPV, ASGV, ACLSV, ApMV, AHVd and ASSVd. Evolutionary distances were computed using the Maximum Composite Likelihood method, and the evolutionary tree was constructed using the Neighbor-Joining method in MEGA11. Bootstrap scores, derived from 1000 replicates, are displayed at each node. CCGaV, citrus concave gum-associated virus; ARWV-1, apple rubbery wood virus 1; ApNMV, apple necrotic mosaic virus; ASPV, apple stem pitting virus; ASGV, apple stem grooving virus; ACLSV, apple mosaic virus; AHVd, apple hammerhead virus; ASSVd, apple scar skin viroid; CiVA, citrus virus A; ARWV-2, apple rubbery wood virus 2; PNRSV, prunus necrotic ringspot virus, PCMV, peach chlorotic mottle virus; PcmV, peach mosaic virus; GLVd, grapevine latent viroid; RdRp, RNA dependent RNA polymerase

exhibited a close genetic relationship with ApNMV previously reported in apples from India (MN832844) (Fig. 4c). The ASPV sequences identified in 14 HTS datasets were associated with various isolates of ASPV of apple from Brazil, Canada, China, Germany, and Japan (Fig. 4d). The ASGV sequences detected in 11 HTS datasets shared genetic similarities with different isolates of ASGV reported in apple from Brazil, Canada, Germany, Japan, and South Korea (Fig. 4e). Among the 10 HTS datasets containing ACLSV, nine were linked to various isolates of ACLSV of apple from Canada, China, Germany, India, Japan, and South Korea. Interestingly, one HTS dataset contained an ACLSV variant that shared

genetic resemblance with ACLSV reported in cherry and *Prunus avium* from Australia, Canada, and the Czech Republic (Fig. 4f). Moreover, ApMV identified in this study clustered together with ApMV isolate (KY965061) infecting apple in Canada (Fig. 4g).

Thirteen AHVd genome sequences identified in this study exhibited close genetic similarities with various isolates of AHVd from apples reported in Belgium, Brazil, Canada, China, India and Italy (Fig. 4h). The AHVd sequences (OR763053) identified from Shimla (Himachal Pradesh) clustered together with AHVd reported in Jammu and Kashmir grouped with different AHVd isolates from

Belgium, Brazil, Canada, India and Italy (Fig. 4h). The AHVd (OR889802) reported from the Oregon Spur cultivar of apple exhibited closest similarities to the AHVd previously reported from the same cultivar in Jammu and Kashmir, India. In addition, the phylogenetic analysis of two ASSVd genome sequences in this study revealed a close genetic relationship with ASSVd (MZ202348) in apple from Italy (Fig. 4i).

Confirmation of HTS-identified viruses using RT-PCR

RT-PCR following Sanger sequencing of cloned viral fragments was performed to confirm the presence of virus contigs detected through HTS (Table S6). RT-PCR results showed the presence of CCGaV and ApNMV in

all 20 apple samples (A3-A23) collected from Jammu and Kashmir (Fig. 5 and Table S6). ARWV-1, ASPV, ASGV and ACLSV were detected in 18, 18, 17 and 10 apple samples of Jammu and Kashmir, respectively (Fig. 5). Moreover, RT-PCR results of 10 apple samples obtained from Himachal Pradesh indicated the presence of CCGaV, ApNMV, ASPV, ASGV and ACLSV, each at varying frequencies (Fig. S3).

The cloned fragments of CCGaV, ApNMV, ARWV-1, ASPV, ASGV and ACLSV shared maximum of 99.8%, 96.94%, 99.3%, 98.42%, 99.75% and 98.4% sequence identities with CCGaV (MK940540), ApNMV (MN832844), ARWV-1 (MF062138), ASPV (KF735118), ASGV (OK338684) and ACLSV (OK340219), respectively.

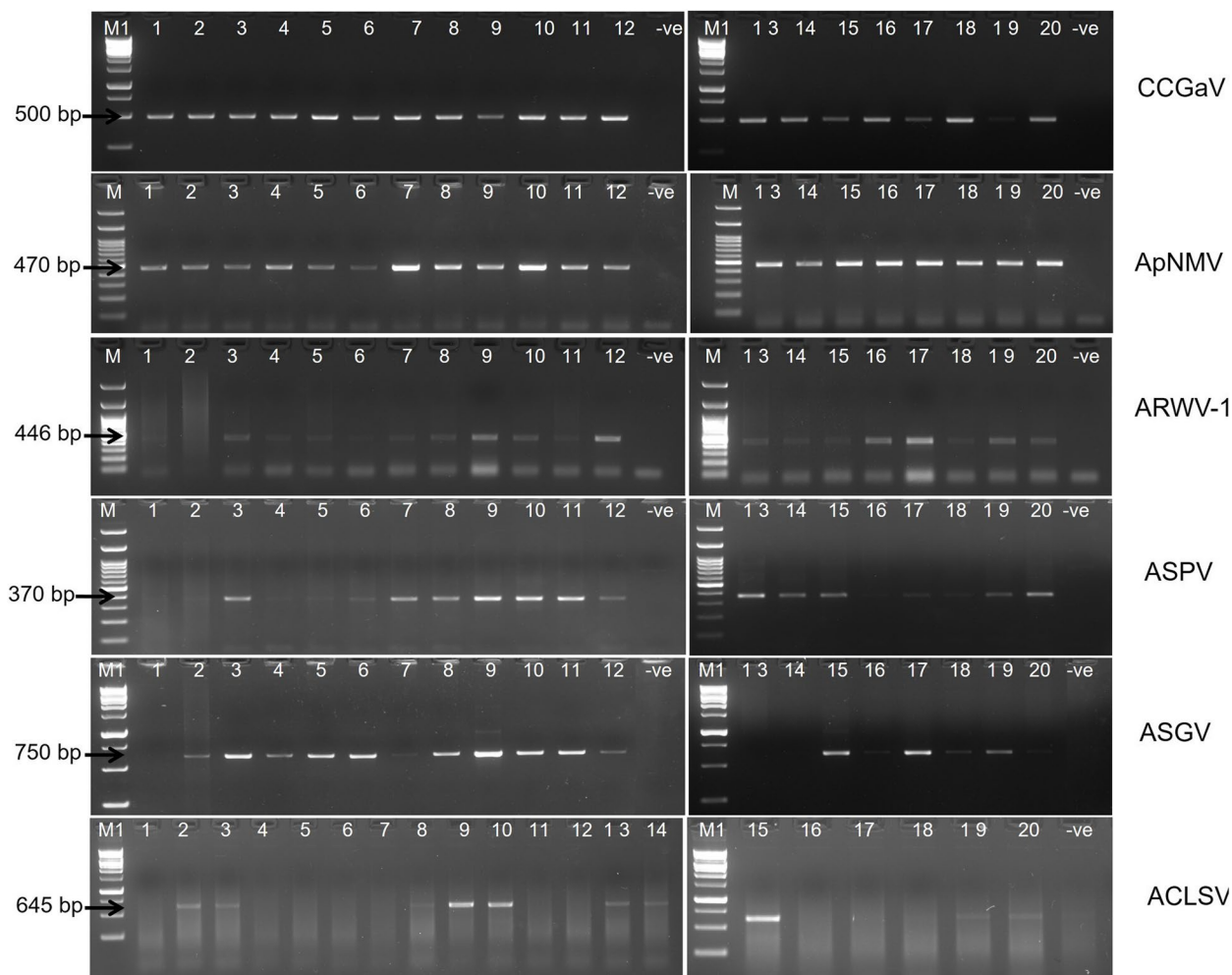


Fig. 5 RT-PCR results showing amplification of partial genomic region of RdRp of CCGaV, replicase of ApNMV, and CP of ARWV-1, ASPV, ASGV and ACLSV from 20 apple samples namely A4 to A23 (lane 1–20) collected from Jammu and Kashmir, using specific primer pairs mentioned in Table S7. Lane M, 100 bp DNA ladder; lane M1: 1 kb DNA ladder, -ve, negative control. CCGaV, citrus concave gum-associated virus; ApNMV, apple necrotic mosaic virus; ARWV-1, apple rubbery wood virus 1; ASPV, apple stem pitting virus; ASGV, apple stem grooving virus; ACLSV, apple chlorotic leaf spot virus; RdRp, RNA dependent RNA polymerase; CP, coat protein

Discussion

Comparing de novo assemblers involves assessing their performance and characteristics in reconstructing accurate and complete genomic sequences from raw sequencing data. This evaluation is crucial for choosing the most suitable assembler for a particular dataset. Numerous studies have compared metagenomic assemblies from bacterial, human and animal sources, highlighting the substantial influence of assembler selection on downstream analysis and the accuracy of reconstructed metagenomes [13–17]. This observation is equally pertinent to the plant viral metagenomes, emphasizing the critical significance of precise and comprehensive assemblies. The assembly step is particularly crucial in virome studies, which face numerous challenges. In this study, we conducted a performance comparison of the Trinity, SPAdes and MEGAHIT assemblers in the context of plant viral metagenomics utilizing 23 apple HTS datasets. Trinity specializes in de novo transcriptome assembly, while SPAdes excel in assembling sequences from both single-cell and multi-cell datasets. MEGAHIT, by contrast, is distinguished by its exceptional speed and memory efficiency, specifically optimized for short-read metagenomic assembly [18–20]. Our comparative analysis provided insights into the strengths and weaknesses of each assembler in handling plant viral metagenomic data, thereby guiding the selection of the most suitable tool for future studies in this domain.

Fragmented assemblies of individual genomes within microbial communities pose obstacles to subsequent analysis and constrain the inferences that can be drawn from metagenomic data, including assessments of taxonomic and functional profiles [21]. To address these challenges, we evaluated several metrics, including percentage genome coverage, genome recovery rates, the generation of larger continuous alignments via single contigs, and N50 values for each assembler. Our analysis revealed that MEGAHIT demonstrated superior performance in several key areas. It produced the longest viral contigs, recovered a higher fraction of the viral genome, achieved a greater genome recovery rate, and obtained higher N50 values compared to Trinity and SPAdes. Additionally, MEGAHIT captured greater viral richness, as indicated by copy number and FPKM, and identified novel SNVs more effectively. These findings are consistent with observations from the SARS-CoV-2 genome assembly study, where metagenomic assemblers (MEGAHIT and metaSPAdes) outperformed other assemblers [15]. By providing a detailed comparison of these metrics, our study highlights the efficacy of MEGAHIT in the context of plant viral metagenomics, suggesting its potential as the assembler of choice for similar datasets in future research.

In our analysis of HTS datasets, we detected SNVs, providing insights into the mutation frequency within virus contigs. We found unique SNVs in apple HTS samples using SPAdes and MEGAHIT. This parallels with the observations in the SARS-CoV-2 genome assembly, where MEGAHIT and metaSPAdes also detected unique SNVs [15]. The presence of SNVs and short InDels varies among assemblers, potentially linked to the assembler's 'aggressiveness' [22]. Variations introduced by assemblers can affect downstream comparative genomics, such as pan-genome comparisons and phylogenetic tree construction using de novo genome assemblies [15].

The N50 value represents the shortest contig length required to cover 50% of the genome. We observed considerable disparities in N50 values among the different assemblers. MEGAHIT consistently exhibited the highest N50 values in comparison to Trinity and SPAdes across all samples. MEGAHIT also assembled longer single contigs in terms of length compared to SPAdes and Trinity. However, our study demonstrated that relying on a single de novo assembler results in an inability to assemble the complete sequences of all viruses or viroids present in the sample.

Apple is a perennial fruit crop of significant economic importance. It is susceptible to various pathogens, with more than twenty viruses and viroids known to infect apple trees worldwide [1]. In this study, the near-complete genome of ASPV, ASGV, ACLSV, ApNMV, ApMV, CCGaV, ARWV-1, ARWV-2, ASSVd and AHVd were obtained and among them CCGaV, ARWV-1, and ARWV-2 have been identified in apple for the first time in India. Our study provides a comprehensive virome profile compared to a previous study done in apple in India [23]. ASGV, ASPV, and ACLSV are considered latent infections, often not causing visible symptoms in most commercially cultivated apple varieties [24, 25]. These viruses lack known natural vectors for transmission, and their spread within apple orchards or nurseries primarily occurs through grafting. Since apple trees are frequently propagated vegetatively through techniques like cuttings and grafting, this can facilitate the transmission and accumulation of viruses [1, 26, 27]. Early detection and eradication of viruses in nurseries and orchards are essential measures to halt their propagation and ensure the production of virus-free planting material. In our HTS analysis, all the apple cultivars exhibiting either mosaic symptoms or those that were asymptomatic showed the presence of viruses and/or viroids. Specifically, ApNMV was detected in apple cultivars displaying mosaic symptoms, such as Golden Delicious (A1), Red Fuji (A2), Royal Delicious (A3), Red Gravenstein (A4), Gala Mast (A5), Spartan (A7), Ambri (A11), Oregon Spur (A15), and American Apirouge (A19). However, Golden

Delicious (A17) displaying mosaic symptoms did not contain ApNMV contigs, while asymptomatic apple cultivars like Golden Delicious (A18), Oregon Spur (A14), and a nursery apple cultivar, Coe Red Fuji (A20), showed the presence of ApNMV contigs. This indicates the importance of HTS-based detection of viruses in plants with or without symptoms.

CCGaV is being reported in apple from India for the first time. It was first reported from Italy in citrus trees and belongs to the genus *Coguvirus* of the family *Phenuiviridae*. It has a bipartite genome consisting of one negative-strand RNA (RNA1) and one ambisense RNA (RNA2). RNA1 encodes the RdRp, while RNA2 encodes the movement protein (MP) and the nucleocapsid protein (NP) [28]. Recently, CCGaV has been identified in apple trees from China, Italy and the USA [29–31]. ARWV-1 and ARWV-2 are members of the genus *Rubodvirus* in the family *Phenuiviridae* and have a tripartite single-stranded negative-sense RNA genome with L, M, and S segments encoding RdRp, MP, and NP, respectively. Both are associated with rubbery wood disease in apple trees. ARWV-1 was first reported in 2018 from Germany and USA and was later identified in apple trees from China [32, 33]. ARWV-2 has been found in apple trees exhibiting decline disease in the United States, as well as in symptomless trees from a global apple collection, and apple trees cultivated in China [31, 34].

Various diagnostic methods have been employed to detect viruses linked to diverse crop diseases. Common methods for virus detection include serological techniques like enzyme-linked immunosorbent assay, molecular methods such as PCR and RT-PCR, isothermal approaches like recombinase polymerase amplification, and the application of CRISPR/Cas12a [35–39]. However, these techniques necessitate the use of polyclonal/monoclonal antibodies and prior knowledge of the targeted virus sequences. Since apple trees are perennial and can be susceptible to multiple pathogens simultaneously, there is a need for a diagnostic approach capable of detecting all potential pathogens in a single analysis. To address these challenges, HTS technology emerges as a valuable tool for the detection of both known and novel viruses. Our study revealed the comprehensive virome profile of apple and identified new viruses, including CCGaV, ARWV-1, and ARWV-2, which were not previously reported in India.

Finally, the confirmation of the identified viruses was achieved through RT-PCR and Sanger sequencing, and with minor exceptions, the results aligned with those inferred from the HTS datasets. RT-PCR-based assays showed a clear dominance of CCGaV and ApNMV, present in almost every sample tested through RT-PCR followed by ARWV-1, ASPV, ASGV and ACLSV. In some

cases, we detected additional viruses through RT-PCR that were not identified using HTS. This discrepancy could be attributed to differences in sensitivity in the two detection techniques [40]. Despite several attempts, we could not confirm the presence of ARWV-2, ApMV, AHVd and ASSVd in the apple sample using RT-PCR. While HTS is generally assumed to be more sensitive than RT-PCR, our study showed that RT-PCR can uncover additional viruses not found in HTS datasets. An earlier study showed that HTS was less sensitive than real-time RT-PCR for the detection of human respiratory viral pathogens including bocavirus and enterovirus [41]. The advantages of HTS over RT-PCR are its ability to provide immediate virus-typing information and identify unknown viruses. RT-PCR is suitable for targeted gene analysis and detecting very low-abundance targets, while HTS is ideal for comprehensive virome analysis.

Conclusions

In conclusion, our analysis of 23 apple HTS datasets using three de novo assemblers demonstrated that MEGAHIT outperformed Trinity and SPAdes in handling plant viral metagenomic data, producing longer viral contigs, higher genome recovery rates, and superior N50 values. The HTS analysis revealed the identification of near-complete genome sequences of eight viruses and two viroids, with CCGaV, ARWV-1, and ARWV-2 identified in apple for the first time in India. Most of the reported viruses were confirmed through RT-PCR and Sanger sequencing. RT-PCR screening results showed CCGaV and ApNMV dominance, followed by ARWV-1, ASPV, ASGV and ACLSV. Virus accumulation studies indicated that ApNMV exhibited the highest copy numbers, followed by ARWV-1 and CCGaV. The phylogenetic analysis showed that the viruses identified in the present study are closely related to viruses found in other countries. Study confirmed the predominance of CCGaV and ApNMV, both of which were present in all samples, while other viruses were detected in varying numbers among the samples tested. Our study emphasizes the pivotal role of assembly methods in constructing apple virus genomes from Illumina reads and their influence on reconstructing nearly complete viral genomes. This study assists in choosing the best assembler for viral genome assembly in perennial fruit crops.

Methods

Sample collection

Leaves from the symptomatic apple (*M. domestica*) plants showing mosaic and necrotic symptoms, as well as asymptomatic apple plants, were collected from the experimental field of ICAR-Central Institute of Temperate Horticulture (CITH; 33.98427° N, 74.80209° E),

Sher-e-Kashmir University of Agricultural Sciences and Technology (SKUAST, 34.14941° N, 74.88407° E), Srinagar, Jammu and Kashmir, a nursery in Kulgam district of Jammu and Kashmir, India and ICAR-Indian Agricultural Research Institute Regional Station, Dhanda (31.10589° N, 77.11737° E), Shimla, Himachal Pradesh, India during May 2022 (Table 1).

RNA extraction, library preparation and RNA sequencing

Total RNA was extracted from 10 symptomatic and 13 asymptomatic plant leaves of 15 apple varieties, comprising of both exotic and indigenous types, as well as one rootstock, using a Spectrum™ Plant Total RNA Kit (Sigma, USA) according to the manufacturer’s guidelines. The quality of RNA was checked using a NanoDrop™ One Microvolume UV–Vis Spectrophotometer (Thermo Scientific, USA). The integrity and quality of the RNA were assessed using an Agilent 4150 TapeStation System (Agilent, USA) and a Qubit 4 Fluorometer (Invitrogen, USA). The rRNA depletion was done using Pan-Plant riboPOOL kit (siTOOLS BIOTECH, Germany) as per the manufacturer’s protocol. RNA with an RNA integrity number greater than 5 and a DV200 score above 85% was subjected to library construction. RNA-seq library preparation was done using a NebNext® Ultra™ II RNA Library Prep Kit for Illumina® (NEB, USA), following the manufacturer’s protocol. Additionally, the library’s quality and

quantity were evaluated using an Agilent 4150 TapeStation and a Qubit 4 Fluorometer (Invitrogen, USA). High-quality total RNA-Seq libraries were subjected to paired-end sequencing (2×150 bp) on a NovaSeq 6000 Sequencing System V1.5 (Illumina, USA). The details of the experimental studies and bioinformatics pipelines are outlined in Fig. 6.

De novo assembly, annotation and comparison of de novo assemblers for the reconstruction of near-complete genome sequences of viruses and viroids

The raw paired-end data (FastQ files) from each sample underwent quality assessment and preprocessing using fastp (version 0.20.1) [42]. We used Bowtie2 (version 2.4.5) to align the clean and filtered sequencing reads to the *M. domestica* genome (ASM211411v1) [43]. For de novo assembly of the host’s unaligned sequence reads, Trinity (version 2.8.5), SPAdes (version 1.3.3), and MEGAHIT (version 1.2.9) were used with default parameters [18–20] (Fig. 6). To annotate the assembled contigs, nonhuman virus sequences were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>) on October 29, 2022. For identification of potential protein coding genes in virus contigs, open reading frames (ORFs) finder (<https://www.ncbi.nlm.nih.gov/orffinder/>) was used. Virus and viroid contigs assembled through different de novo assemblers in 23 apple HTS datasets

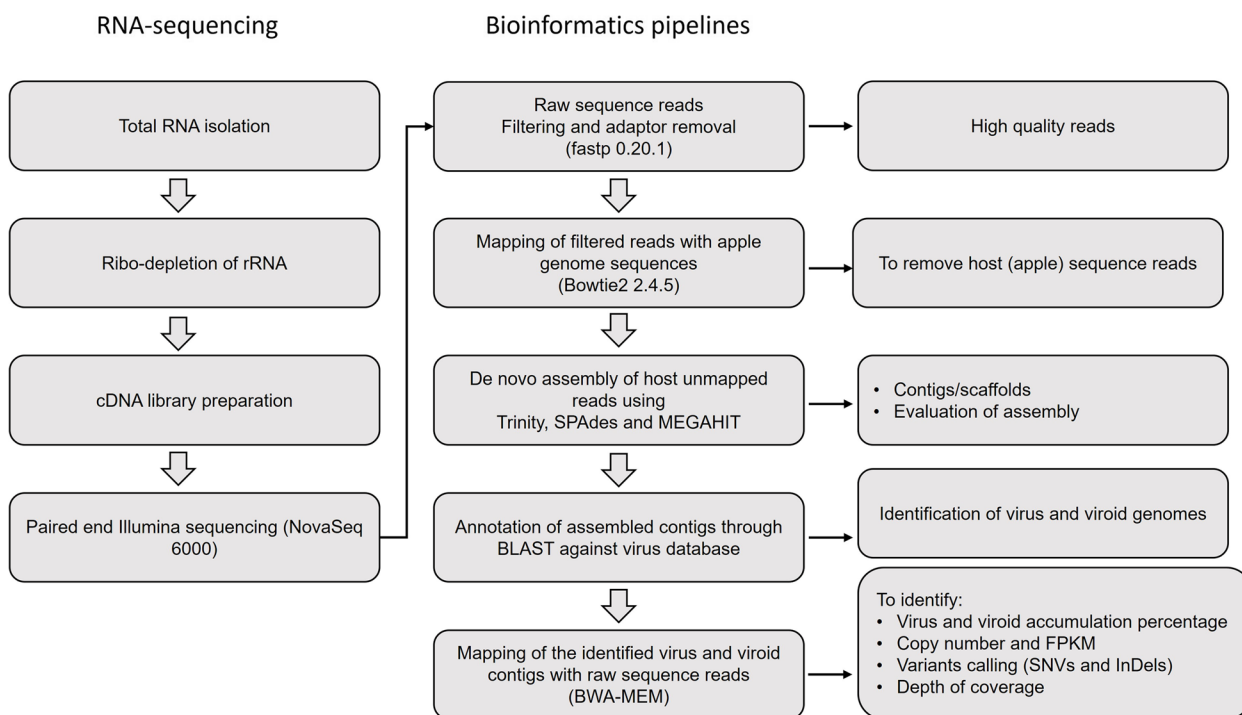


Fig. 6 Illustration of RNA sequencing and bioinformatic workflows for apple virome analysis

were compared. De novo assembled contigs representing near-complete genome sequences of CCGaV, ApNMV, ApMV, ASPV, ASGV, ACLSV, ARWV-1, AHVd, and ASSVd were submitted to GenBank.

Comparing de novo assemblers for genome mapping, variant calling, depth of coverage and virus richness of assembled contigs

For the alignment of sequence data to the reference virus and viroid genomes, the BWA-MEM algorithm was employed [44]. Host-unmapped, filtered reads from each individual sample were aligned against the reference sequences of viruses and viroids. Virus and viroid specific reads were then extracted from the resulting sequence alignment map (SAM) files. These reads were used to determine virus/viroid accumulation, depth of coverage, copy number and FPKM. SAMtools was used to process the aligned reads, generating an mpileup file for variants calling. VarScan was used to identify SNVs and insertions and deletions (InDels) from SAMtools mpileup data [45]. Variant calls (SNVs and InDels) were extracted from the VCF file using bcftools with default parameters.

After obtaining the read count for each virus/viroid from the SAM file, the percentage of reads per virus/viroid was computed. This was achieved by dividing the reads attributed to each specific virus/viroid by the total reads encompassing all viruses and viroids, and then multiplying by 100. The number of virus-associated reads was multiplied by 150 bp and divided by the length of each virus' genome to get the copy number. The calculation of FPKM involved multiplying the read count for each individual virus/viroid by 10^9 and dividing the result by the total accumulated virus and viroid reads, which were further multiplied by the length of the respective virus/viroid [46].

The impact of different assemblers on virus copy numbers, FPKM values, and SNVs was assessed and compared. The performance of de novo assemblers was evaluated for genome coverage percentage, contig alignment length, and variant calling using the assembled virus and viroid contigs as reference sequences.

Phylogenetic analysis

We selected RNA dependent RNA polymerase (RdRp) gene of CCGaV, ARWV-1 and ASPV, replicase gene of ApNMV, ApMV and ACLSV, and polyprotein gene of ASGV, and complete nucleotide sequences of AHVd and ASSVd for the construction of phylogenetic dendrogram. These nucleotide sequences were retrieved from GenBank separately. These sequences along with the virus and viroid contigs of present study underwent multiple sequence alignment using ClustalW.

Subsequently, phylogenetic trees were constructed using Neighbor-Joining method with a bootstrap test (1000 replicates) and evolutionary distances were calculated using the Maximum Composite Likelihood method, expressed as base substitutions per site, in Molecular Evolutionary Genetic Analysis version 11 [47].

Confirmation of viruses

To confirm the presence of HTS identified virus contigs in the apple plant samples, RT-PCR was performed using the same RNA employed for HTS and additional apple leaf samples. Total RNA was isolated from apple plant leaves using a Spectrum™ Plant Total RNA Kit (Sigma, USA) following manufacturer's instructions. RNA was checked on non-denaturing agarose gel electrophoresis. The quality and quantity of RNA were assessed using the NanoDrop™ One Microvolume UV–Vis Spectrophotometer (Thermo Scientific, USA). For first-strand cDNA synthesis, the FIREScript® RT cDNA synthesis kit (Solis BioDyne, Estonia) was employed, following the manufacturer's protocol. Subsequently, PCR was carried out using 10 μM of each primer set (Table S7), 50 ng cDNA, and 1X DreamTaq PCR Master Mix (Thermo Scientific, USA). The PCR cycle conditions comprised an initial denaturation at 94 °C for 4 min followed by 35 cycles of denaturation at 94 °C for 45 s, annealing at 50–61 °C for 45 s (depending on the specific primer pairs) and extension at 72 °C for 30–45 s (depending on specific primer pairs) and a final extension at 72 °C for 10 min (Table S7). The amplified DNA was subjected to electrophoresis on a 1.6% agarose gel, stained with EtBr, and visualized using a gel documentation system (Bio-Rad, USA). To further analyze, the PCR amplicons were cloned into pGEM®-T Easy Vector (Promega, USA), and 5, 8, 4, 7, 8 and 3 bacterial colonies of different samples having genomic fragments of CCGaV, ApNMV, ARWV-1, ASPV, ASGV and ACLSV were sequenced using Sanger sequencing technology. The sequences of the cloned fragments were then checked using BLASTn.

Abbreviations

ACLSV	Apple chlorotic leaf spot virus
AHVd	Apple hammerhead viroid
ApMV	Apple mosaic virus
ApNMV	Apple necrotic mosaic virus
ARWV-1	Apple rubbery wood virus 1
ARWV-2	Apple rubbery wood virus 2
ASGV	Apple stem grooving virus
ASPV	Apple stem pitting virus
ASSVd	Apple scar skin viroid
CCGaV	Citrus concave gum-associated virus
FPKM	Fragments per kilobase of transcript per million mapped reads
HTS	High-throughput sequencing
RT-PCR	Reverse transcription polymerase chain reaction
SNVs	Single nucleotide variants

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10968-x>.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary material 7.
Supplementary material 8.

Acknowledgements

We are thankful to Head, ICAR-Indian Agricultural Research Institute Regional Station, Amartara Cottage, Shimla, Director, ICAR-Central Institute of Temperate Horticulture, Srinagar, and Dr. Mehraj D. Shah, Department of Plant Pathology, Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir, Shalimar, Srinagar for providing apple leaf samples.

Authors' contributions

VKB and SKS conceived and designed the study. ZAK, VKB, SUN and SW performed the field work and collected the samples. ZAK, DD, PT and MYS performed the lab experiments. ZAK, DD, PT, SKS, NG, MRP, MKV and VKB analysed the data. ZAK wrote the manuscript. SKS and VKB reviewed and edited the manuscript. All authors have read and approved the manuscript.

Funding

The research work was supported by Indian Council of Agricultural Research under National Professor Scheme to Professor V.K. Baranwal (Ag.Edn.F.No. /27/01/NP/2022-HRD).

Data availability

The datasets generated and/or analysed during the current study are available in NCBI SRA. The weblink of the SRA are: <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1025925?reviewer=cts6cjna4ed5vthlodae8stmpj>, <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1032713?reviewer=ai5vtie2tIs39qfdghafjtpmp1>, <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1034804?reviewer=qlhspj2lnk74fqa2v4q8otola>, <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1034824?reviewer=kf7fk2e0q8djl5ttgkplh9204q>.

Declarations

Ethics approval and consent to participate

There are no research using human participants or animals in the article.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Advanced Centre for Plant Virology, Division of Plant Pathology, ICAR-Indian Agricultural Research Institute, New Delhi 110012, India. ²ICAR-Central Institute of Temperate Horticulture, Srinagar 191132, India. ³ICAR-Indian Agricultural Research Institute, Regional Station, Shimla, Himachal Pradesh 171004, India. ⁴Current Address: United States Department of Agriculture, Agricultural Research Service, Northern Crop Science Laboratory, Fargo, ND 58102, USA.

Received: 30 May 2024 Accepted: 28 October 2024
Published online: 08 November 2024

References

- Xiao H, Hao W, Storoschuk G, MacDonald JL, Sanfaçon H. Characterizing the virome of apple orchards affected by rapid decline in the Okanagan and Similkameen Valleys of British Columbia (Canada). *Pathogens*. 2022;11:1231.
- Sedlák J, Příbylová J, Koloňuk I, Špak J, Lenz O, Semerák M. Elimination of *Solanum nigrum* ilarvirus 1 and apple hammerhead viroid from apple cultivars using antivirals ribavirin, rimantadine, and zidovudine. *Viruses*. 2023;15:1684.
- Chen S, Ye T, Hao L, Chen H, Wang S, Fan Z, et al. Infection of apple by apple stem grooving virus leads to extensive alterations in gene expression patterns but no disease symptoms. *PLoS One*. 2014;9:e95239–e95239.
- Malandraki I, Beris D, Isaioglou I, Olmos A, Varveri C, Vassilakos N. Simultaneous detection of three pome fruit tree viruses by one-step multiplex quantitative RT-PCR. *PLoS One*. 2017;12:e0180877.
- Walia Y, Dhir S, Ram R, Zaidi AA, Hallan V. Identification of the herbaceous host range of apple scar skin viroid and analysis of its progeny variants. *Plant Pathol*. 2014;63:684–90.
- Xing F, Robe BL, Zhang Z, Wang H, Li S. Genomic analysis, sequence diversity, and occurrence of apple necrotic mosaic virus, a novel ilarvirus associated with mosaic disease of apple trees in China. *Plant Dis*. 2018;102:1841–7.
- Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. *Cell Mol Life Sci*. 2016;73:4433–48.
- Duffy S. Why are RNA virus mutation rates so damn high? *PLoS Biol*. 2018;16:e3000003.
- Schneider WL, Roossinck MJ. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *J Virol*. 2001;75:6566–71.
- Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog*. 2010;6:e1001005.
- Dufault-Thompson K, Jiang X. Applications of de Bruijn graphs in microbiome research. *iMeta*. 2022;1:e4.
- Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform*. 2019;20:1140–50.
- Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome*. 2019;7:1–15.
- Greenwald WW, Klitgord N, Seguritan Y, Yooseph S, Venter JC, Garner C, et al. Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC Genomics*. 2017;18:1–11.
- Islam R, Raju RS, Tasnim N, Shihab IH, Bhuiyan MA, Araf Y, et al. Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Brief Bioinform*. 2021;22:bbab102.
- Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep*. 2016;6:19233.
- White DJ, Wang J, Hall RJ. Assessing the impact of assemblers on virus detection in a de novo metagenomic analysis pipeline. *J Comput Biol*. 2017;24:874–81.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–77.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*. 2011;29:644.
- Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31:1674–6.
- Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS One*. 2011;6:e21400.
- Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 2012;22:557–67.
- Nabi SU, Baranwal VK, Rao GP, Mansoor S, Vladulescu C, Raja WH, et al. High-throughput RNA sequencing of mosaic infected and non-infected apple (*Malus domestica* Borkh.) cultivars: from detection to the reconstruction of whole genome of viruses and viroid. *Plants*. 2022;11:675.
- Hadidi A, Barba M, Candresse T, Jelkmann W. Virus and virus-like diseases of pome and stone fruits. *Am Phytopath Society*; 2011.
- Massart S, Jijakli MH, Kummert J. CHAPTER 7: Apple stem grooving virus. In: *Virus and virus-like diseases of pome and stone fruits*. *Am Phytopath Society*; 2011. p. 29–33. <https://doi.org/10.1094/9780890545010>.

26. Umer M, Liu J, You H, Xu C, Dong K, Luo N, et al. Genomic, morphological and biological traits of the viruses infecting major fruit trees. *Viruses*. 2019;11:515.
27. Yaegashi H, Yoshikawa N, Candresse T. CHAPTER 4: Apple chlorotic leaf spot virus in pome fruits. In: *Virus and virus-like diseases of pome and stone fruits*. Am Phytopath Society; 2011. p. 17–21. <https://doi.org/10.1094/9780890545010>.
28. 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.
29. Liu Z, Dong Z, Zhan B, Li S. Characterization of an isolate of citrus concave gum-associated virus from apples in China and development of an RT-RPA assay for the rapid detection of the virus. *Plants*. 2021;10:2239.
30. Minutolo M, Cinque M, Chiumenti M, Di Serio F, Alioto D, Navarro B. Identification and characterization of citrus concave gum-associated virus infecting citrus and apple trees by serological, molecular and high-throughput sequencing approaches. *Plants*. 2021;10:2390.
31. Wright AA, Szostek SA, Beaver-Kanuya E, Harper SJ. Diversity of three bunya-like viruses infecting apple. *Arch Virol*. 2018;163:3339–43.
32. Hu GJ, Dong YF, Zhang ZP, Fan XD, Ren F, Lu XK. First report of apple rubbery wood virus 1 in apple in China. *Plant Dis*. 2021;105:3770.
33. Rott ME, Kesanakurti P, Berwarth C, Rast H, Boyes I, Phelan J, et al. Discovery of negative-sense RNA viruses in trees infected with apple rubbery wood disease by next-generation sequencing. *Plant Dis*. 2018;102:1254–63.
34. Hu GJ, Dong YF, Zhang ZP, Fan XD, Ren F, Lu XK. First report of apple rubbery wood virus 2 infection of apples in China. *Plant Dis*. 2021;105:519.
35. Jeong J, Ju H, Noh J. A review of detection methods for the plant viruses. *Res Plant Dis*. 2014;20:173–81.
36. Jiao J, Kong K, Han J, Song S, Bai T, Song C, et al. Field detection of multiple RNA viruses/viroids in apple using a CRISPR/Cas12a-based visual assay. *Plant Biotechnol J*. 2021;19:394–405.
37. Kim N-Y, Oh J, Lee S-H, Kim H, Moon JS, Jeong R-D. Rapid and specific detection of apple stem grooving virus by reverse transcription-recombinase polymerase amplification. *Plant Pathol J*. 2018;34:575.
38. Komorowska B, Malinowski T, Michalczyk L. Evaluation of several RT-PCR primer pairs for the detection of Apple stem pitting virus. *J Virol Methods*. 2010;168:242–7.
39. Menzel W, Jelkmann W, Maiss E. Detection of four apple viruses by multiplex RT-PCR assays with coamplification of plant mRNA as internal control. *J Virol Methods*. 2002;99:81–92.
40. Fowkes AR, McGreig S, Pufal H, Duffy S, Howard B, Adams IP, et al. Integrating high throughput sequencing into survey design reveals turnip yellows virus and soybean dwarf virus in pea (*Pisum sativum*) in the United Kingdom. *Viruses*. 2021;13:2530.
41. Prachayangprecha S, Schapendonk CME, Koopmans MP, Osterhaus ADME, Schürch AC, Pas SD, et al. Exploring the potential of next-generation sequencing in detection of respiratory viruses. *J Clin Microbiol*. 2014;52:3722–30.
42. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34:i884–90.
43. Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min*. 2015;8:1–7.
44. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*. 2013.
45. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568–76.
46. Khan ZA, Diksha D, Thapa P, Maillem YS, Sharma SK, Gupta N, et al. Genome analysis of viruses of *Phenuiviridae*, *Betaflexiviridae* and *Bromoviridae*, and apple scar skin viroid in pear by high-throughput sequencing revealing host expansion of a rubodivirus and an ilarvirus. *Physiol Mol Plant Pathol*. 2023;102196. <https://doi.org/10.1016/j.pmp.2023.102196>.
47. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol*. 2021;38:3022–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.