MBP-FRIDAY-651

June 14, 2024

Does Long-Read Sequencing Technology Produce Superior Viral Genome Assemblies?

ATCC®
Credible leads to Incredible®

Corina L. Tabron, MS; Nikhita P. Puthuveetil, MS; Jade L. Kirkland, MS; Kaitlyn Gaffney, MS; Noah Wax, MS; James Duncan, BS; Robert Marlow, BS; Stephen King, MS; Ana Fernandes, BS; John Bagnoli, BS; Briana Benton, BS; Jonathan L. Jacobs, PhD | ATCC, Manassas, VA 20110

Background

Authenticated and traceable genomic data is vital for reproducible science, whether for preclinical studies, drug discovery, host-virus interaction, therapeutic development, or countless other applications. While the genomes for many of ATCC®'s viruses are available in public databases, as previously shown, these reference genomes published by third parties are often error-prone, incomplete, or generated using a variety of methods with a lack of supporting metadata, making downstream analysis challenging.[1] To address this problem, we developed reproducible next-generation sequencing and genome assembly workflows to produce viral assemblies for over three hundred viruses within our diverse collection. Our viral assemblies are generally produced with only short-read sequencing technology. Here, we set out to determine if we could produce higher-quality assemblies using long-read sequencing technology, starting with DNA viruses.

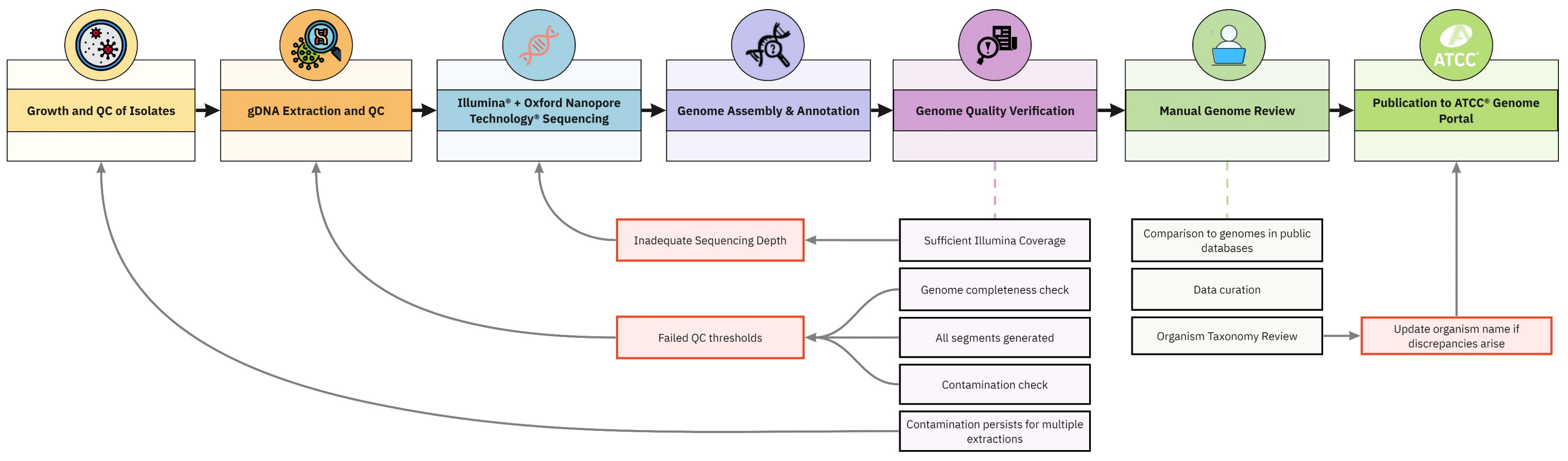

Growth and QC of Isolates
gDNA Extraction and QC
Illumina® + Oxford Nanopore Technology® Sequencing
Genome Assembly & Annotation
Genome Quality Verification
Manual Genome Review
Publication to ATCC® Genome Portal
ATCC
Inadequate Sequencing Depth
Failed QC thresholds
Sufficient Illumina Coverage
Genome completeness check
All segments generated
Contamination check
Contamination persists for multiple extractions
Comparison to genomes in public databases
Data curation
Organism Taxonomy Review
Update organism name if discrepancies arise

Figure 1: Standard ATCC® Genome Portal publication workflow.

Explore ATCC® Virology Genomes

Methods

Table 1: Preliminary extraction kit selection. We currently generate viral assemblies from starting material extracted with the QIAGEN® QIAamp® Viral RNA Mini Kit (catalog no. 52904, QIAGEN®, MD, USA). This method supports our minimum DNA concentration requirement of 1 ng for library preparation on Illumina® platforms. However, to integrate long-read sequencing into our viral pipeline, a method that yields concentrations ≥5 ng was required to be in line with our internal Oxford Nanopore Technologies® library preparation protocols. This was achieved with the QIAGEN® MinElute® Media Kit (catalog no. 57414, QIAGEN®, MD, USA).

ATCC®	Organism	Qubit® (ng/µL)	
		QIAamp® Viral Mini Kit	QIAamp® MinElute® Media Kit
VR-1090™	Human adenovirus D	3.67	20.7
VR-197™	Simian adenovirus	1.30	9.05
VR-3393™	Human herpesvirus 2	1.19	7.29
VR-1491™	Human gammaherpesvirus 4	3.22	25.2

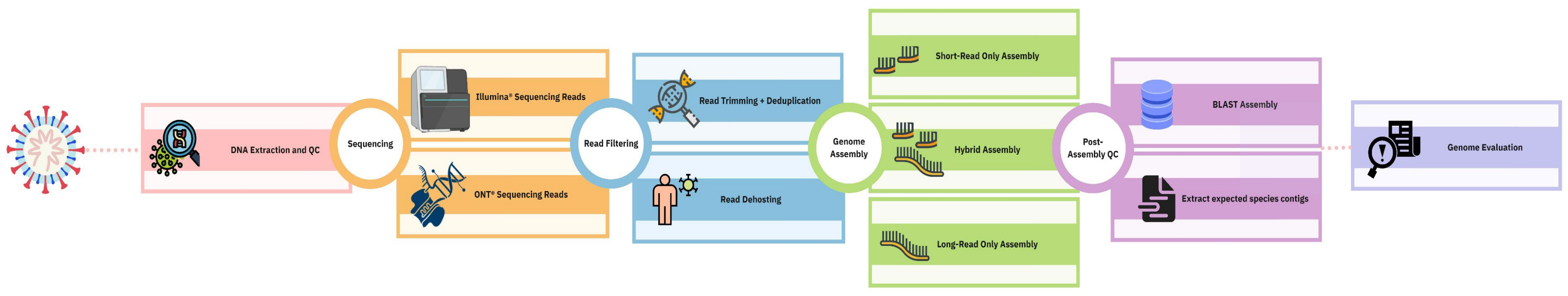


Figure 2: Workflow used to produce viral assemblies. Isolated nucleic acids from viruses (see Table 1) were divided and processed for short-read sequencing on the Illumina® MiSeq® and for long-read sequencing on the Oxford Nanopore® GridION®. Unicycler[2] and Flye[3] were used to generate individual de novo assemblies from the ONT® and Illumina® reads, as well as hybrid de novo assemblies from data produced on both platforms. Final assemblies were then evaluated for quality.

References

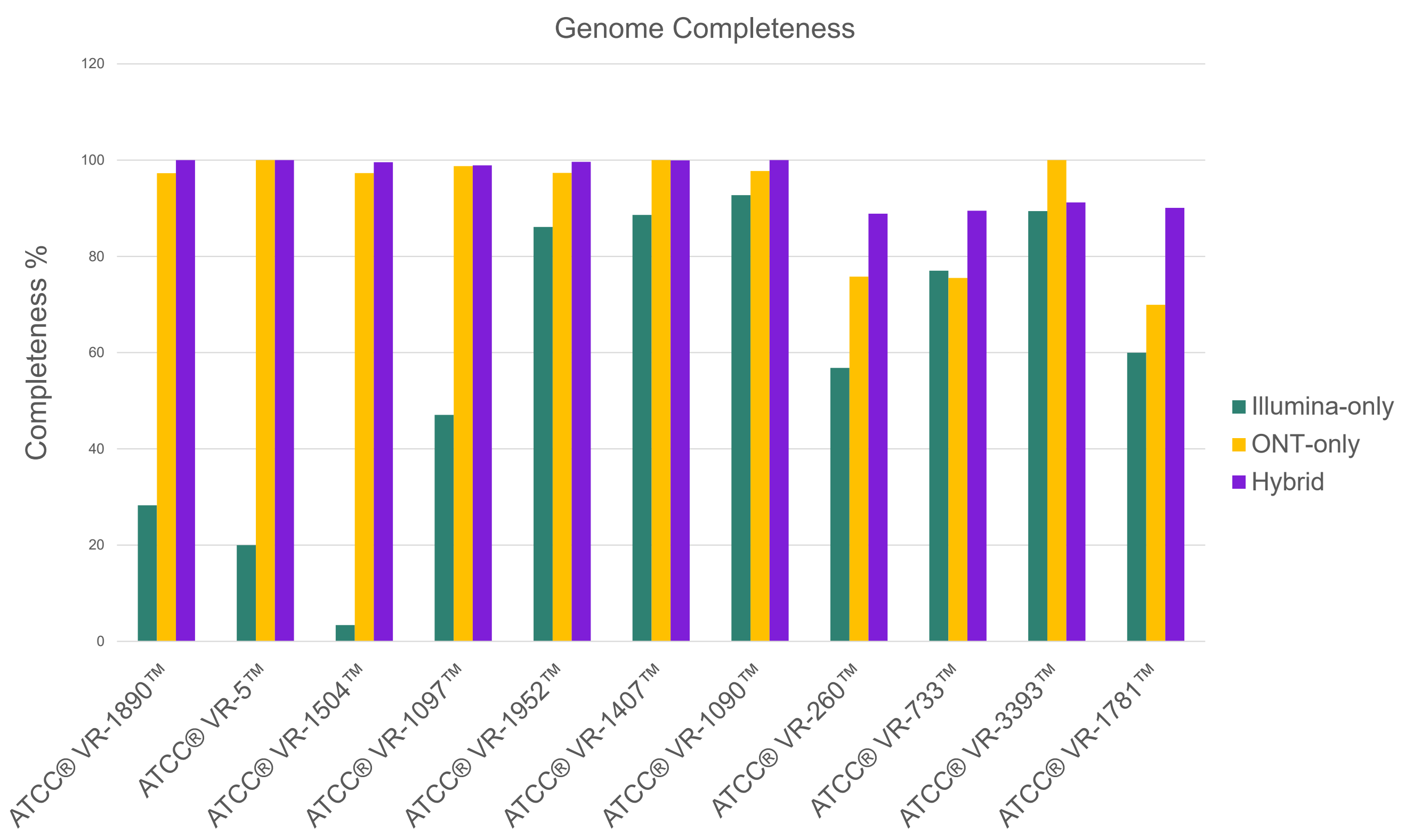
1. Yarmosh DA, et al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. 7(3): e0007722, 2022.
2. Wick RR, et al. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595, 2017.
3. Kolmogorov M, et al. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnol 37: 540-546, 2019.
4. Wick RR, et al. Bandage: interactive visualisation of de novo genome assemblies. Bioinformatics, 31(20): 3350-3352, 2015.

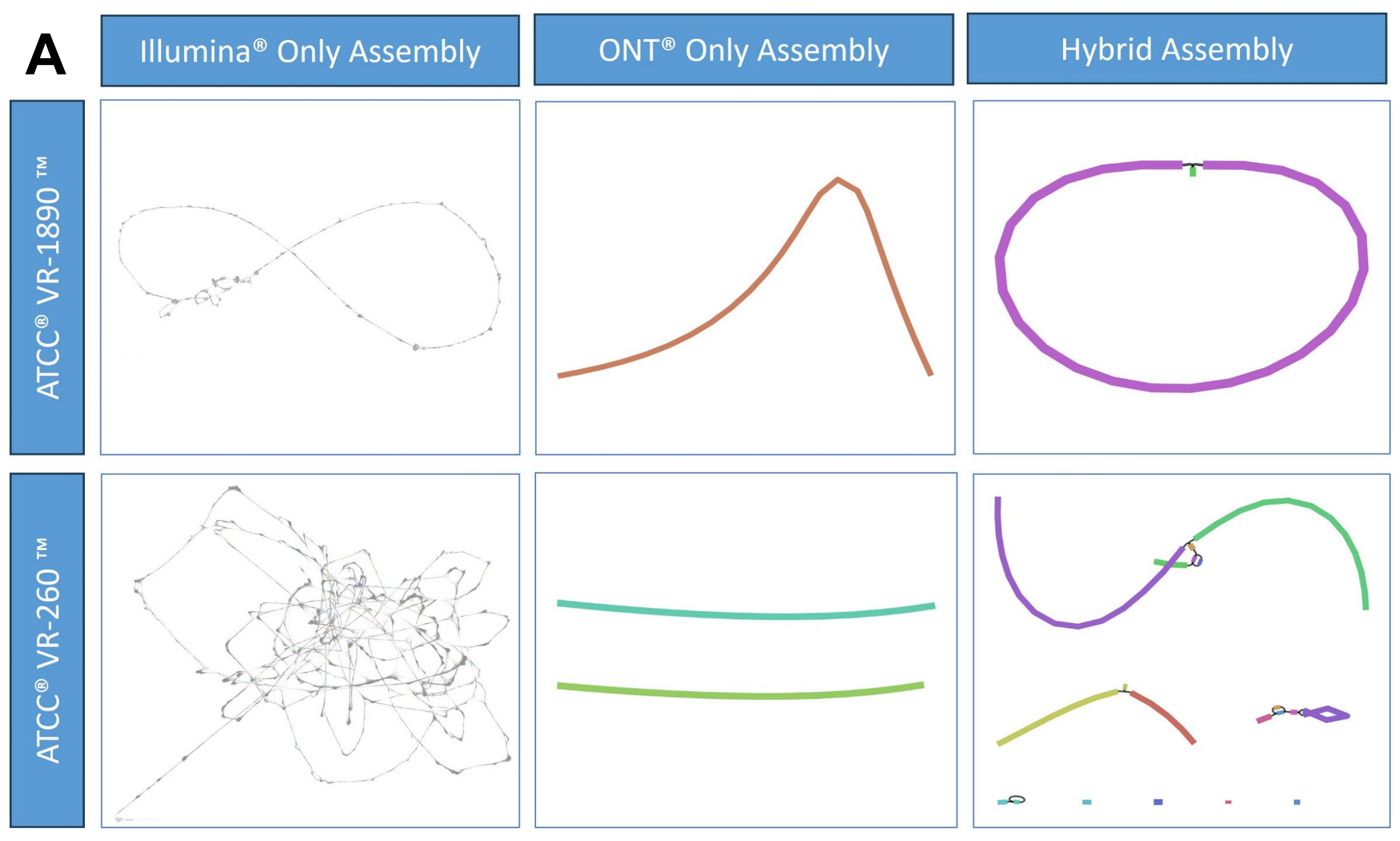
Table 2: Genome assembly length and contiguous DNA fragment counts were recorded for 11 DNA viruses produced by each assembly method.

ATCC®	Virus	Assembly Length			Contigs		
		Illumina®	ONT®	Hybrid	Illumina®	ONT®	Hybrid
VR-1890™	Human adenovirus 1	11,076	33,987	35,783	15	1	1
VR-5™	Human adenovirus 5	7,481	35,872	35,725	5	1	1
VR-1504™	Human adenovirus 10	1,163	33,937	34,783	2	1	1
VR-1097™	Human adenovirus 20	18,385	34,753	35,729	23	1	2
VR-1952™	Human adenovirus 33	34,196	34,006	34,811	29	1	1
VR-1407™	Human adenovirus 49	34,623	35,122	34,915	26	1	1
VR-1090™	Human adenovirus D	34,477	34,395	35,020	23	1	1
VR-260™	Human herpesvirus 1	111,863	115,246	134,147	102	2	6
VR-733™	Human herpesvirus 1	129,692	132,133	134,672	78	3	4
VR-3393™	Human herpesvirus 2	137,564	168,060	138,296	7	1	7
VR-1781™	Human herpesvirus 2	125,735	136,351	138,133	124	1	4

Results

Figure 3: CheckV genome completeness comparison across three different assembly methods for 11 viral strains.





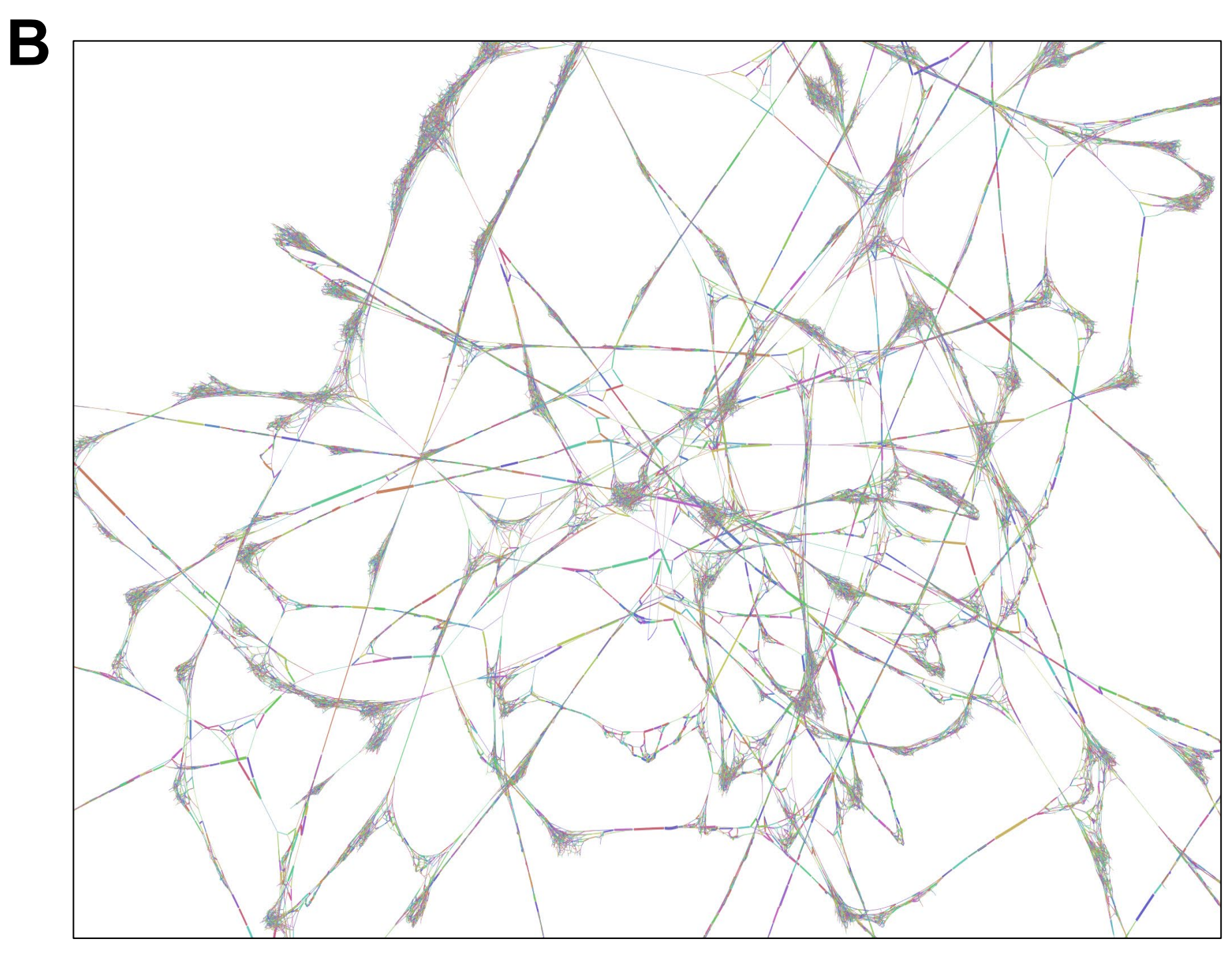


Figure 4: Graphical Fragment Assembly (GFA) files were created by each assembler during the assembly process. The GFAs were then visualized using Bandage.[4] (A) The GFAs for ATCC® VR-1890™ and ATCC® VR-260™ for each assembly method are shown above. Overall, the graphs created using only Illumina® sequencing reads indicate poor assemblies. The assemblies improve with the inclusion of ONT® reads or with only assembling with ONT® reads, as these graphs are much less complex and have fewer edges. (B) Close-up image of the tangled Illumina®-only assembly for ATCC® VR-260™.

Conclusions

- Our findings show that the inclusion of long-read sequencing technology can improve genome assembly quality for DNA viruses as several strains assembled with long-reads exhibited a higher genome completeness and a total assembly length closer to what is expected for the strain.
- With the inclusion of long reads, assemblies had a drastic decrease in contig counts compared to their short-read only counterparts.
- Based on these results, utilizing both short-read and long-read technology in our viral assembly pipelines generates higher-quality assemblies.
- Further work will be conducted to optimize workflows dedicated to processing RNA viruses using long-read sequencing technology.

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 **Phone:** 800.638.6597 **Email:** sales@atcc.org **Web:** www.atcc.org