Generalizing Herpesvirus Genome Assembly: A Novel Bioinformatics Approach

Nikhita Puthuveetil, MS; Corina Tabron, MS; Joseph Petrone, PhD; Amy Reese, MS; David Yarmosh, MS; John Bagnoli, BS; and Jonathan Jacobs, PhD
ATCC, Manassas, VA 20110

Herpesvirus Genomes

Herpesviruses are highly prevalent DNA viruses that can cause recurring, lifelong infections in humans and other mammals. The genomes of these viruses are notoriously difficult to assemble. With short-read only approaches, resulting assemblies are often fragmented. Genomes assembled using long-read only approaches are more complete, but they often fail to capture the inverted terminal repeat (ITR) regions. Publicly available genomes tend to use a manual approach that involves laboriously curating each contig and specialized sequencing protocols. Due to manual intervention, these assemblies are more contiguous; however, for high-throughput labs, such methods cannot be easily employed. Here, we present our herpesvirus pipeline (HvP) and compare it against three other custom ATCC® assembly methods.

Figure 1: Characteristics of Herpesvirus genomes. Due to their large genome size (125-240 kb in length), extremely high GC-content, and numerous terminal and internal repeat regions, herpesviruses can be a challenge to accurately assemble.

Pipeline Design

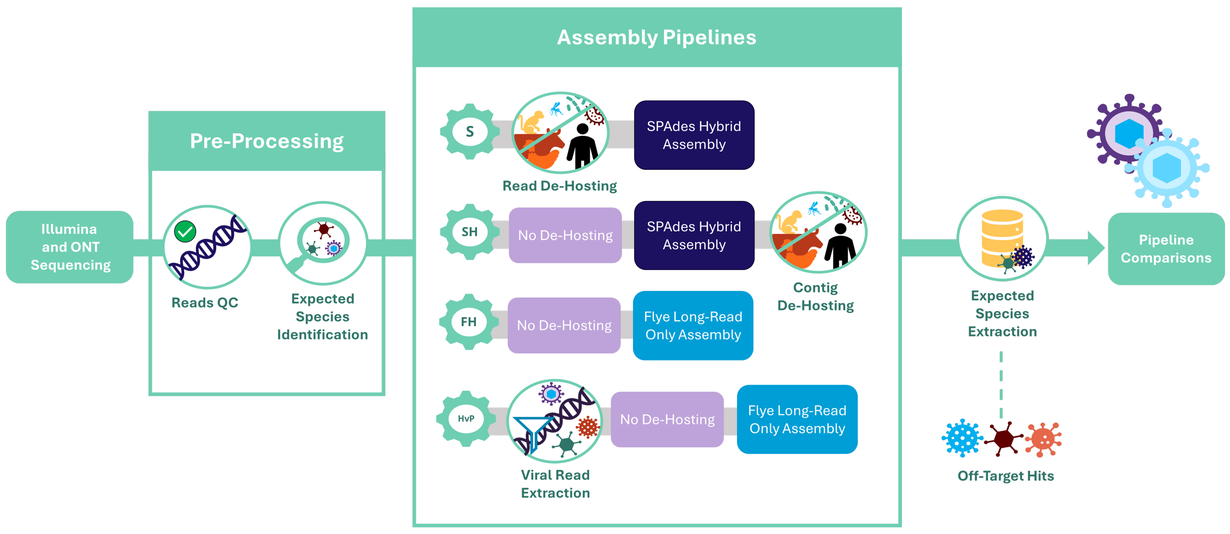

Figure 2: Assembly pipelines. All samples passed through the same read pre-processing steps. Processed reads were then each put through four different assembly methods: **Standard (S), Standard with Host (SH), Flye with Host (FH),** and **Herpesvirus Pipeline (HvP)**. All generated assemblies then passed through the same post processing steps and were compared against each other.

Table 1: Eleven different species of herpesviruses of various families and lengths were used to compare all four pipelines.

Virus Name	Subfamily	Genome Size (bp)	Available Genomes in NCBI
Canid herpesvirus 1	α	125,171	11
Caprine herpesvirus	α	134,617	2
Bovine herpesvirus 1	α	134,896	50
Felid herpesvirus 1	α	135,797	55
Herpes simplex virus-1	α	152,222	108
Herpes simplex virus-2	α	156,293	34
Human herpesvirus 6B	β	162,114	5
Epstein-Barr virus	γ	171,823	586
Gallid herpesvirus 2	α	177,874	66
Murine Cytomegalovirus	β	230,408	19
Human Cytomegalovirus	β	235,646	348

Genomes Comparison Results

Table 2: Results summary.

ATCC® Catalog Number	Name	Expected Length	Generated Assembly Length by Pipeline			
			Standard	Standard w/ Host	Flye w/ Host	HvP
VR-1789™	Herpes simplex virus-1	152,222	108,188	43,409	208,594	152,624
VR-1779™	Herpes simplex virus-2	156,293	132,788	112,206	N/A	154,684
VR-864™	Bovine herpesvirus 1	134,896	124,788	123,832	136,598	136,567
VR-1785™	Canid herpesvirus 1	125,171	115,729	165,214	147,054	128,508
VR-814™	Felid herpesvirus 1	135,797	124,552	123,989	149,814	136,084
VR-1576™	Gallid herpesvirus 2	135,797	124,552	123,989	149,814	136,084

Generally, assemblies produced with Flye were closer to the expected species length and assemblies produced with SPAdes were much more fragmented. Compared to other assemblies, the assemblies generated by the HvP pipeline were more consistently complete and in line with the expected species length. The FH pipeline did not generate an assembly for ATCC® VR-1779™.

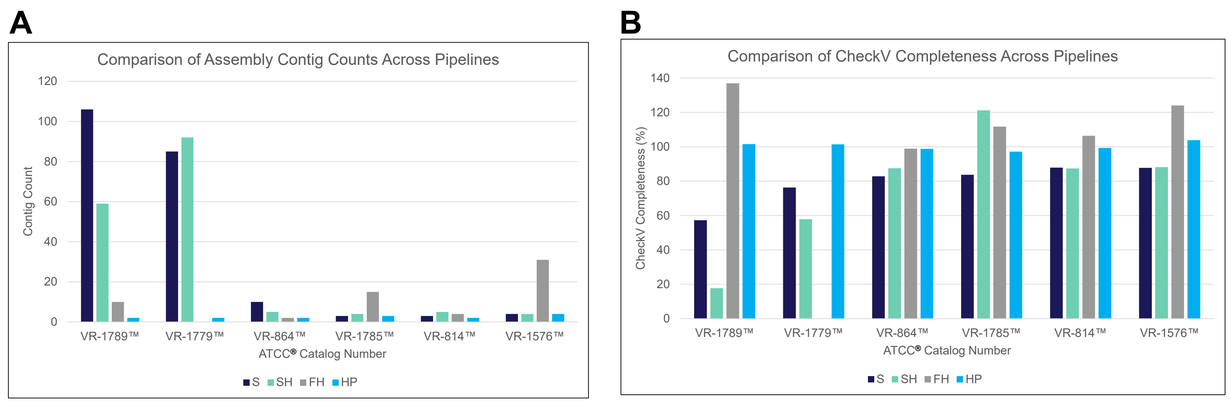

Figure 3: Contig Counts and Completeness Comparisons. Assemblies generated using the HvP or the FH pipeline performed better compared to the other two SPAdes pipelines. Plots (A) and (B) compare the differences in assembly contig counts and CheckV completeness among the four pipelines with HvP generating assemblies with lower contig counts and higher completeness scores.

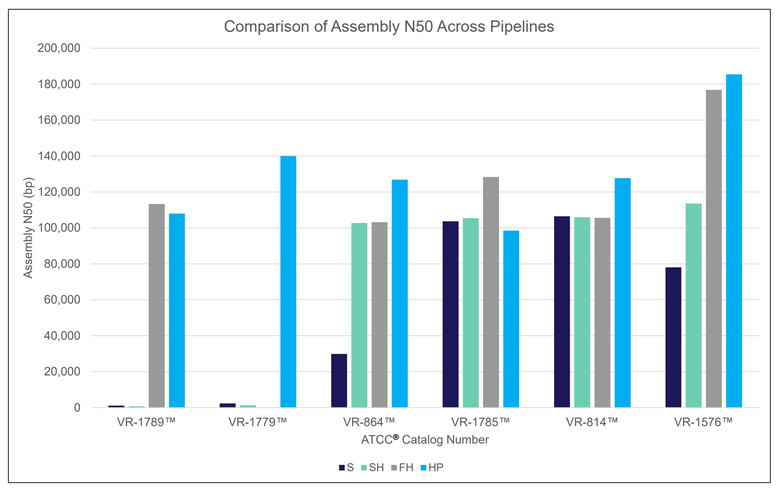

Figure 4: N50 Comparisons. Assemblies generated by the HvP pipeline had higher N50s across samples while assemblies generated with SPAdes did poorly. The FH pipeline also performed well, but not as consistently as the HvP pipeline.

Herpesviruses and the ATCC® Genome Portal

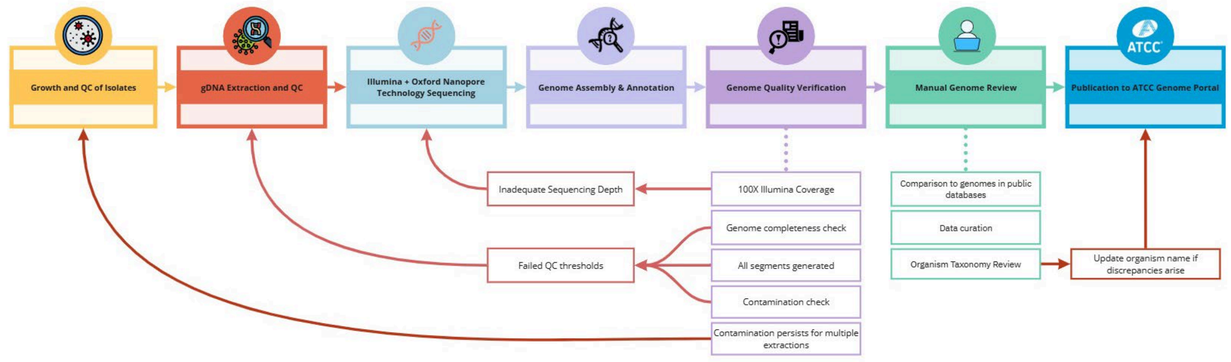

Figure 5: Publication Workflow to ATCC® Genome Portal (AGP). Genomes on the AGP, ATCC®'s highly authenticated genomic database, go through an extensive manual review process prior to publication. Even after publication, assemblies are continuously improved to ensure genomes are of the highest quality. The HvP pipeline was developed in an effort to improve and add additional herpesviruses to the AGP.

SCAN ME

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