PIP-TUES-131
October 15, 2024

Building a Sequence Repository

SCAN ME

ATCC®
Credible leads to Incredible®

Nikhita P. Puthuveetil, MS; John Bagnoli, BS; Joseph R. Petrone, PhD; David A. Yarmosh, MS; Amy L. Reese, MS; Jonathan L. Jacobs, PhD | ATCC, Manassas, VA 20110

Background

In 2019, the ATCC® Genome Portal (AGP) was launched as a part of an initiative to produce high-quality reference genomes for the entirety of the ATCC microbial collection. As of October 2024, ATCC has published 5,000 microbial genomes complete with all supporting metadata. Here, we present the bioinformatics workflow as of January 1, 2024, that standardizes each genomic assembly made available through the AGP. ATCC®'s microbial assembly pipeline is composed of discrete modules for each category of organisms: viruses, bacteria, and microbial eukaryotes. We recognize the challenges in adapting pipelines to accommodate the highly diverse organisms under these broad taxonomic levels. We present these pipelines as a comprehensive solution for generating high-quality genomes in a high-throughput setting.

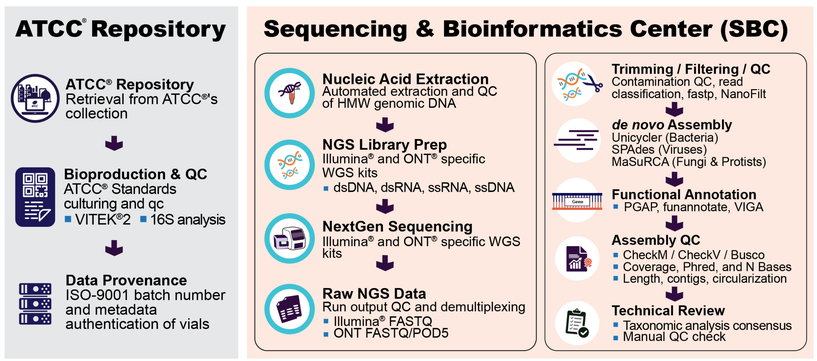

Figure 1: Standard pipeline detailing the process of genome portal publication from sample deposition to the ATCC® collection through final publication to the AGP.

Bacteriology Module

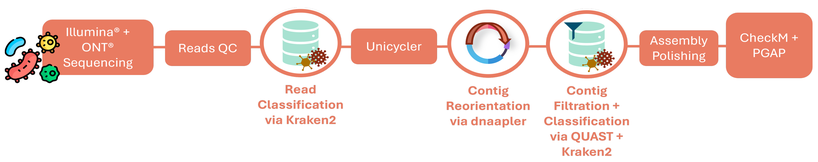

Figure 2: Details on the bacteriology module of ATCC's assembly pipeline. All bacterial samples are sequenced on both sequencing platforms. The subsequent reads are trimmed and filtered, assembled through Unicycler, and then go through final polishing and assembly evaluation steps.

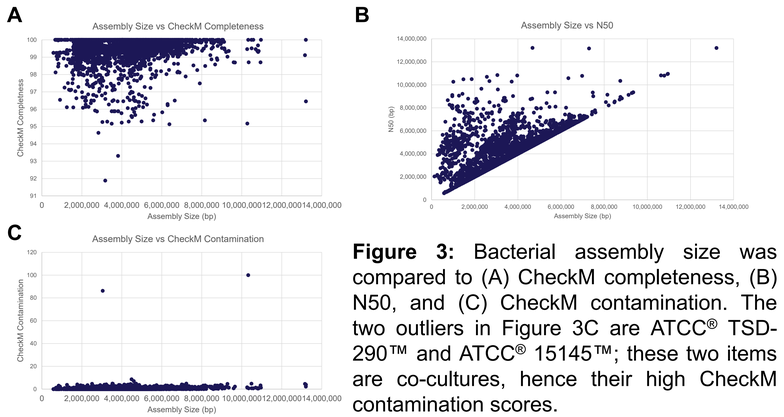

Figure 3: Bacterial assembly size was compared to (A) CheckM completeness, (B) N50, and (C) CheckM contamination. The two outliers in Figure 3C are ATCC® TSD-290™ and ATCC® 15145™; these two items are co-cultures, hence their high CheckM contamination scores.

Virology Module

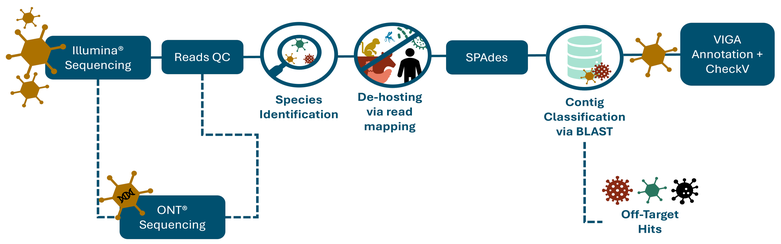

Figure 4: Details of the virology module of ATCC®'s assembly pipeline. All viral samples are sequenced on Illumina®. DNA viruses are also sequenced on ONT®. Notably, samples go through an in-silico de-hosting step where reads that did not map to a custom host multifasta are used as input for the SPAdes assembler.

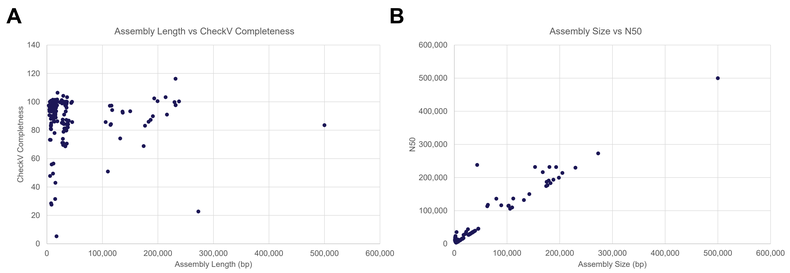

Figure 5: Viral assembly size was compared to (A) to viral completeness and (B) N50. Samples with low viral completeness have been manually evaluated prior to AGP publication.

Microbial Eukaryotes Module

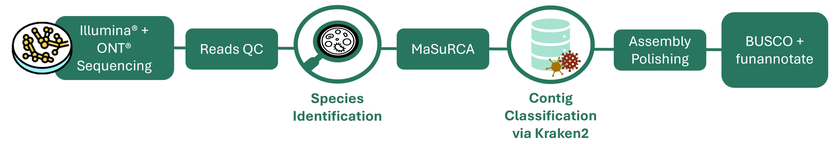

Figure 6: Details of the microbial eukaryote module of ATCC®'s assembly pipeline. All microbial eukaryotic samples are sequenced on both sequencing platforms and are assembled using the MaSuRCA assembler.

"White Glove" Assemblies

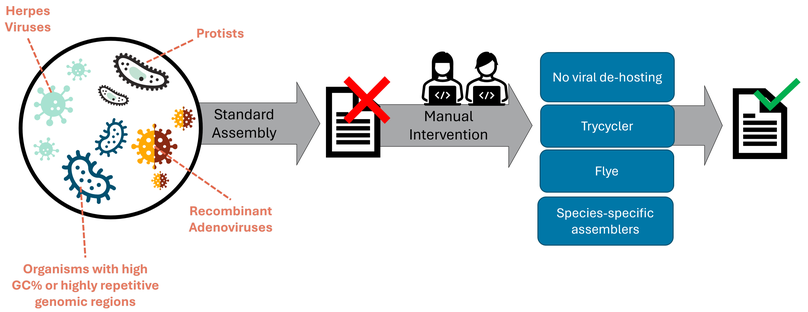

Figure 7: ATCC®'s microbial assembly pipeline is not always able to completely assemble certain organisms like herpesviruses or organisms with highly repetitive genomic regions. Instead, these organisms are manually assembled and evaluated using a variety of techniques.

Conclusions

The design of ATCC®'s microbial assembly pipelines has allowed our team to operate quickly and efficiently to produce the thousands of genomes available through the ATCC® Genome Portal while standardizing the assembly process for reproducible, high-quality genomes. Even so, we have adapted a manual process, or "white glove" approach for problematic genomes that confound our automated pipelines. We are continuing to refine the pipelines into narrower taxonomic categories to further streamline automation. Through these and accompanying detailed metadata, ATCC® can provide the scientific community with traceable, authenticated assemblies from productions from the ATCC® collection.

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