MBP-FRIDAY-655

June 16, 2023

The ATCC® Genome Portal: Expanded Authenticated Microbial Reference Genomes with Data Provenance

ATCC®

Credible leads to Incredible®

Jonathan Jacobs, PhD; Nikhita Puthuveetil, MS; Briana Benton, BS; John Bagnoli, BS; Ana Fernandes, BS; David Yarmosh, MS; Amy L. Reese, MS; Joseph Petrone, PhD; James Duncan, MS; Samuel Greenfield, BS; Corina Tabron, MS; Jade Kirkland; MS; Noah Wax, MS; Robert Marlow, BS; Stephen King, MS; Hannah McConnell; Scott Nguyen, PhD
ATCC, Manassas, VA 20110

Background

The ATCC® Genome Portal is a multi-year initiative aimed at producing high-quality microbial reference genomes representing the entire microbial collection at the American Type Culture Collection (ATCC®). All data is publicly accessible, curated, and traceable to the physical materials in ATCC's collection. As of April 2023, the ATCC® Genome Portal included fully authenticated genome assemblies and annotations for over 2,778 bacterial, 250 viral, 207 fungal, and 4 protist genomes. All sequencing data, assemblies, and annotations were produced in-house at ATCC®. We present our progress in expanding the ATCC Genome Portal.

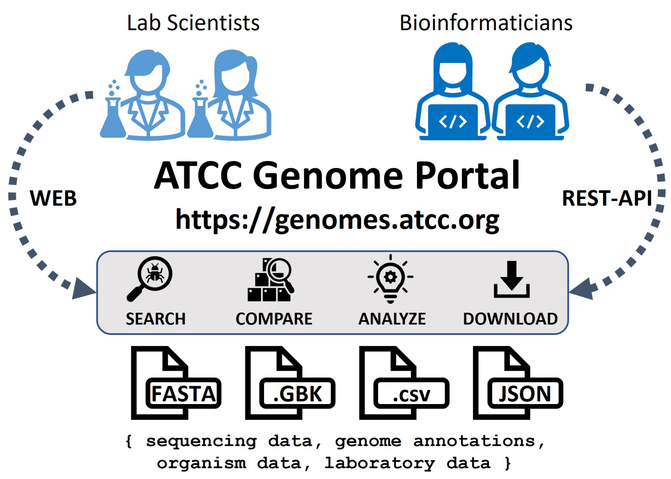

Figure 1: ATCC® Genome Portal functionality. The ATCC® Genome Portal is a freely available resource for research-use purposes and is accessible via the web (https://genomes.atcc.org) or via an authenticated REST-API. Users can download authenticated and traceable genome assemblies, annotations, and metadata.

AGP Publication

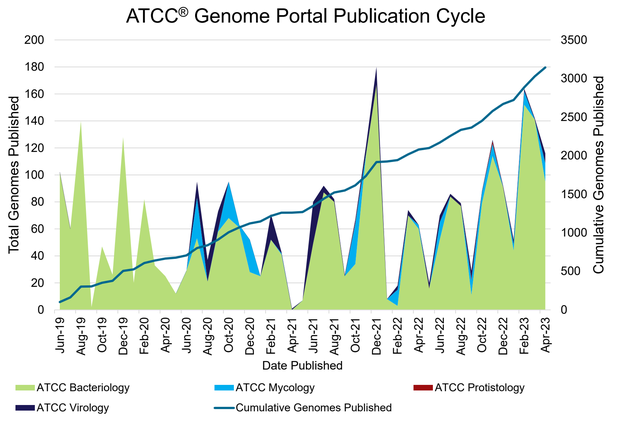

Figure 2: ATCC® Genome Portal Publication Cycle. Genomes are routinely released on a monthly basis to the ATCC® Genome Portal. In 2020, viral and fungal genomes were introduced to the portal. In 2022, protist genomes were introduced. As of April 2023, more than 3,000 genomes are currently available for download.

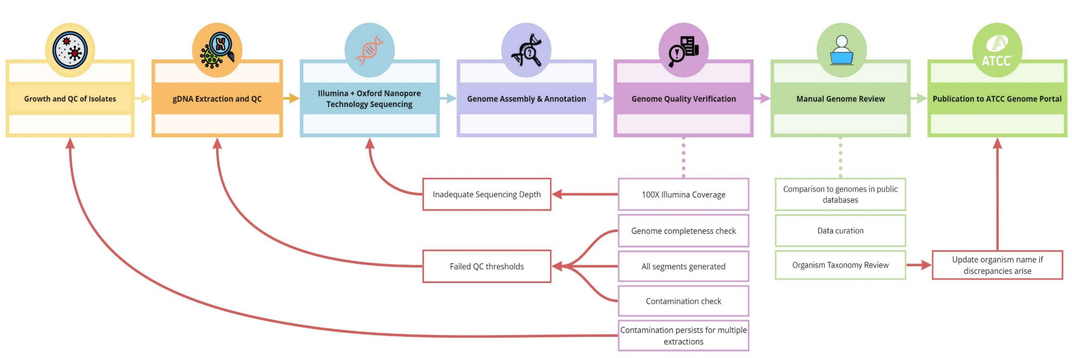

Figure 3: ATCC® Genome Portal Publication Workflow. The genome publication process begins with growth and QC of isolates. Next, gDNA is extracted and then sequenced on two sequencing platforms. The sequencing reads are assembled using kingdom-specific assemblers and the resultant genome is verified and manually curated. If the genomes fails QC, the sample returns to the lab for re-extraction/re-sequencing and is re-assembled. Passing genomes and their annotations are published to ATCC® Genome Portal.

AGP Organism Diversity

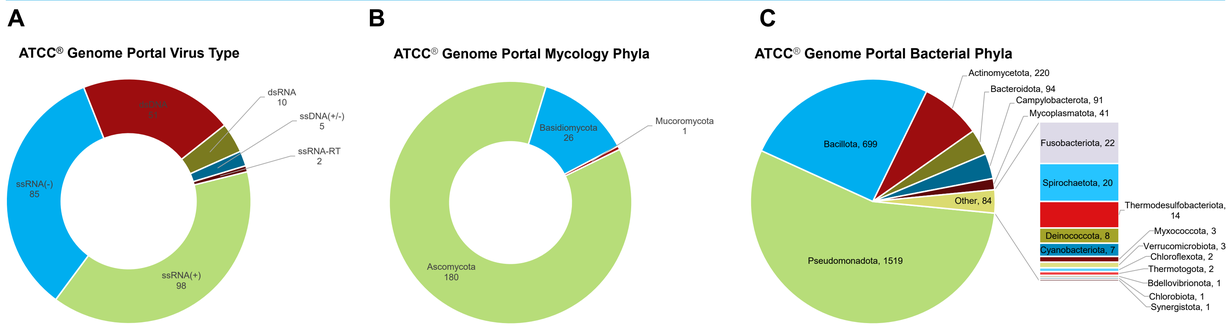

Figure 4: Organism Diversity of ATCC Genomes. Distribution of organisms across genomes published to the ATCC® Genome Portal. (A) Distribution of viral genomes by virus type. (B) Distribution of fungal genomes by phyla. (C) Distribution of bacterial genomes by phyla.

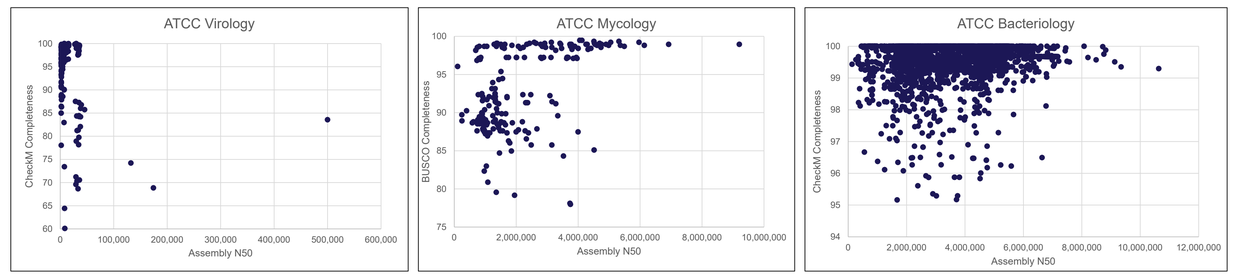

Figure 5: Assembly N50 vs Completeness. Comparison of Assembly N50 to Genome Completeness for each ATCC® Collection. Bacterial and viral assemblies are assessed for completeness using CheckM.[1] Fungal assemblies are analyzed by BUSCO.[2] Certain genomes may still be published despite lower metrics after extensive manual review.

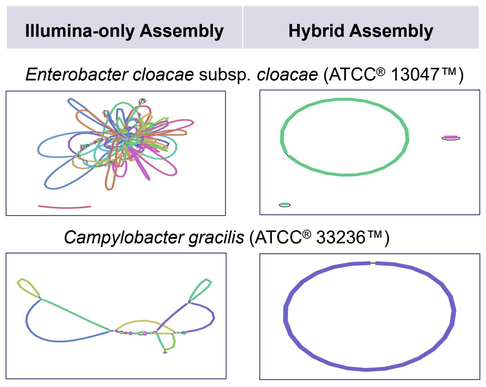

Figure 6: Illumina-only vs Hybrid Assembly. Comparison of genomes assembled with only Illumina short reads or with both Illumina and Oxford Nanopore long reads. Both fungal and bacterial genomes along with select viruses are assembled with a hybrid approach.

Understudied Microbes

Table 1: Understudied Microbes Sequenced in 2023

ATCC® Catalog Number	Name	Genome Portal Assembly Level	Number of Publications	Number of GenBank Assemblies	Number of GenBank Nucleotide entries	Applications	Industries of Interest
51179™	*Pseudodesulfovibrio halophilus*	Complete	22	0	7	Arsenate/sulfate reduction	Biotechnology
31628™	*Tatumella terrea*	Chromosome	16	0	50	Production of Vitamin C precursor, wine contaminant	Nutrition, Wine, Paint
TSD-267™	*Proteiniphilum* sp.	Complete	0	0	44	Unknown	Unknown
700555™	*Borrelia andersonii*	Chromosome	64	0	80	Causes Lyme disease	Infectious Disease
BAA-2643™	*Porphyromonas pogonae*	Complete	8	0	6	Causes infections in central bearded dragons	Infectious Disease
BAA-1452™	*Bartonella durdenii*	Complete	4	0	5	Causes infections in squirrels	Infectious Disease
700995™	*Halomonas maura*	Chromosome	29	0	21	Nanofiber and biofilm generation, increase alfalfa yield	Biotechnology, Nanotechnology, Agriculture
TSD-283™	*Caproiciproducens reaktori*	Chromosome	0	0	0	Unknown	Unknown
TSD-242™	*Denitromonas iodocrescerans*	Chromosome	0	0	1	Iodate reduction	Biodegradation
BAA-173™	*Tsukamurella strandjordii*	Complete	6	0	36	Causes sepsis, potential biofertiliser	Infection Disease, Agriculture
BAA-108™	*Pandoraea* genomospecies 2	Chromosome	0	0	0	Causes infections in humans	Infectious Disease
700638™	*Janibacter hoylei*	Chromosome	18	7	112	Degradation of naphthalene, phenathrene	Biodegradation
43327™	*Pasteurella stomatis*	Chromosome	45	0	14	Causes infections in humans after animal exposure	Infectious Disease
96294™	*Hyphozyma lignicola*	Complete	2	0	7	Trehalose ester production, plant pathogen	Nutrition/Food, Agriculture
66751™	*Hyphozyma variabilis*	Scaffold	2	0	14	Unknown	Unknown
62894™	*Candida saitoana*	Scaffold	28	0	49	Induces resistance in fruit to fungal infection, generates peptides for cosmetics	Biocontrol, Agriculture, Skincare/Cosmetics
MYA-4126™	*Dactylellina leptospora*	Scaffold	3	0	30	Biocontrol agent of parasitic nematodes	Biocontrol

These bacterial and fungal genome assemblies published in 2023 have few to no available assemblies on public databases. Despite the lack of representation and publications, these organisms are quite versatile, having applications of a variety of industries ranging from agriculture to cosmetics and skincare.

Conclusions

Many assemblies for strains represented in public databases are not required to be authenticated, nor even traceable to physical biomaterials in a biorepository or culture collection. Due to traceability issues, a changing landscape of sequencing technologies and bioinformatics methods, and the near absence of requirements for metadata harmonization, the quality and reliability of microbial genomics data in the public domain has steadily declined. This complicates many downstream bioinformatics applications and research outcomes due to unexpected, yet often substantial, discrepancies between the physical strains present in culture collections and the genomes that represent those strains in public databases.[3] The ATCC® Genome Portal is intended to address this gap in data provenance and data quality for ATCC strains.

Learn about the ATCC® Enhanced Authentication Initiative

References

1. Parks D, Skennerton C, Imelfort M. CheckM [Source code] https://github.com/Ecogenomics/CheckM, 2014.
2. Simão F, Waterhouse R, Ionnidis I, Kriventseva E, Zdobnov E. BUSCO [Source Code] http://busco.ezlab.org/, 2015.
3. Yarmosh D, et. al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. mSphere 7(3): e00077-22, 2022.

Special thanks to all the members of the Sequencing & Bioinformatics Center (SBC) team

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