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

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.

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-sequenceing and is re-assembled. Passing genomes and their annotations are published to ATCC® Genome Portal.

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. Fungal assemblies are analyzed by BUSCO. 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

<table>
<thead>
<tr>
<th>ATCC® Catalog</th>
<th>Name</th>
<th>Source</th>
<th>Number of Bases</th>
<th>Number of Bases annotated</th>
<th>Application</th>
<th>Detection of disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAA-2643™</td>
<td>Porphyromonas sp.</td>
<td>Environmental</td>
<td>22,000,000</td>
<td>12,000,000</td>
<td>Infectious Disease</td>
<td>Unknown</td>
</tr>
<tr>
<td>BAA-3178™</td>
<td>Pasteurella sp.</td>
<td>Environmental</td>
<td>24,000,000</td>
<td>14,000,000</td>
<td>Infectious Disease</td>
<td>Unknown</td>
</tr>
<tr>
<td>BAA-1903™</td>
<td>Candida sp.</td>
<td>Environmental</td>
<td>30,000,000</td>
<td>18,000,000</td>
<td>Infectious Disease</td>
<td>Unknown</td>
</tr>
<tr>
<td>BAA-173™</td>
<td>Hyphozyma lignicola</td>
<td>Environmental</td>
<td>5,000,000</td>
<td>5,000,000</td>
<td>Infectious Disease</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

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. The ATCC® Genome Portal is intended to address this gap in data provenance and data quality for ATCC strains.

References


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