The ATCC® Genome Portal

Authenticated Microbial Reference Genomes with Data Provenance

ATCC®

Credible leads to Incredible®

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

Abstract

Publicly available microbial genomes are commonly used in industry research for a variety of purposes. These data, however, are often poor quality or lack critical metadata, which can lead to delays in research progress and increased costs.

In response, we created the ATCC® Genome Portal: a regularly updated database of *de novo* genome assemblies and annotations for microbes held in the ATCC collection. Currently, the ATCC Genome Portal includes 3,118 genome assemblies produced in-house by ATCC from materials sourced directly from our biorepository. The database currently provides references for 2,679 bacteria, 243 viruses, 192 fungi, and 4 protists—including genome assemblies for 1,007 type-strains. The content of the ATCC Genome Portal is updated every month with new genome assemblies, and we aim to reach 10,000 authenticated genomes by 2025. Here, we describe our standardized workflows, data diversity, and recent comparative genomics highlights from the genome portal.

Each bacterium, fungus, and protist is sequenced on both Illumina and Oxford Nanopore platforms; the results of which are used to produce hybrid *de novo* assemblies for each strain. Viruses are currently sequenced using only Illumina sequencing. Each assembly on the ATCC Genome Portal includes metrics for sequencing quality, assembly completeness, genome annotations, and metadata such as antibiotic susceptibility, isolation data, origins, and phenotypic data.

The ATCC Genome Portal and the data contained therein is freely available for research-use and is accessible via the web or via a new REST-API. Please visit https://genomes.atcc.org for details.

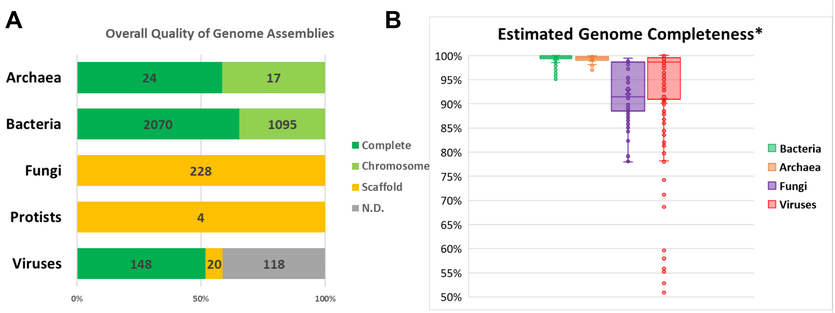


Figure 1: ATCC Genome Portal Assembly Quality and Completeness. (A) Number and quality of assembles in the ATCC Genome Portal according to NCBI's assembly quality descriptors. (B) Box plot of estimated genome completeness (% vertical axis) for each assembly by kingdom.

*Genome "completeness" estimations with CheckM, BUSCO, and the NCBI Viral Annotation Pipeline are dependent on pre-existing databases which, for many of our genomes, no prior genome reference exists. Thus, we expect these estimates to be lower than the actual level of completeness.

References

If you use our data as part of your own research, please site the following references:

1. Benton B et al. ***The ATCC Genome Portal: Microbial Genome Reference Standards with Data Provenance***. *Microbiol Resour Announc* **2021** https://doi.org/10.1128/MRA.00818-21.

2. Yarmosh DA et al. ***Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies***. *mSphere* **2022** https://doi.org/10.1128/msphere.00077-22.

Methods

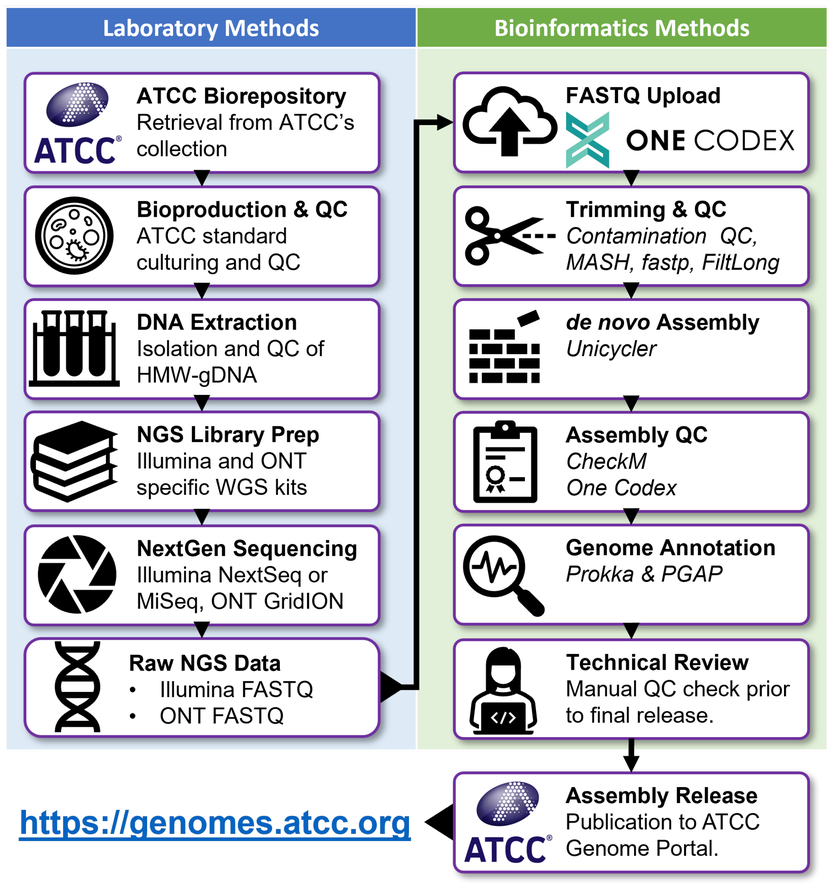


Figure 2: Pipeline for end-to-end genomic data provenance. Source materials were obtained directly from the ATCC® biorepository and tracked through to the final assembly and genome annotation. Upfront culture conditions varied depending on the species cultured, but downstream process steps were performed using standardized protocols for DNA extraction, library prep, sequencing, and bioinformatics. Each pipeline is hosted on One Codex®.

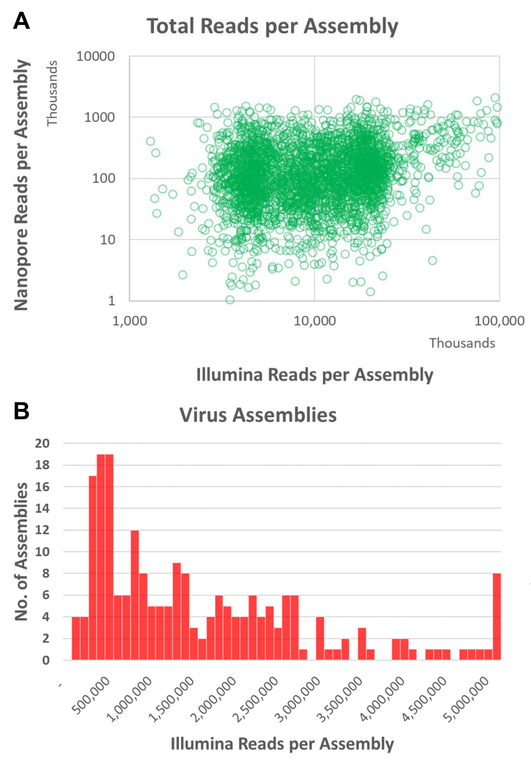
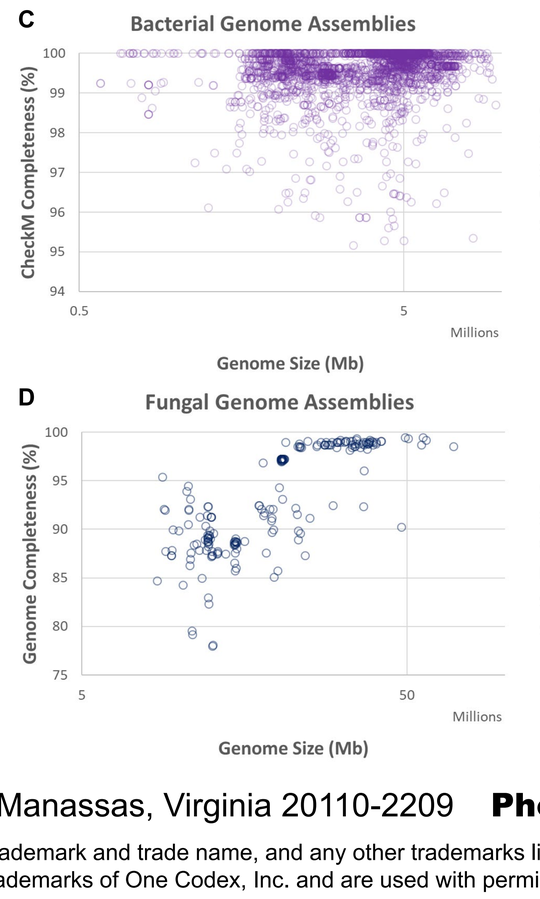


Figure 3: Sequencing depth of Illumina and Nanopore.

(A) Distribution of reads for bacteria, archaea, and fungus assemblies.

(B) Distribution of only Illumina reads for all virus assemblies.



(C) Distribution of bacterial genome sizes and overall genome completeness by CheckM. X axis in Log(10) scale.

(D) Distribution of fungal genome sizes and overall completeness as measured by BUSCO score. X axis in Log(10) scale.

Results

Table 1: Bacteria

Phylum	# Genome Assemblies	Avg. % Completeness	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Actinomycetota	257	99.4	11,716,101	164,393	3,105,392
Bacillota	831	99.4	11,214,339	204,431	2,853,768
Bacteroidota	103	99.5	10,993,728	175,046	3,288,007
Chlorobiota	1	98.9	10,335,140	23,086	2,154,823
Chloroflexota	2	99.1	5,498,763	19,269	3,646,753
Cyanobacteriota	9	99.5	7,544,248	84,031	4,563,488
Deinococcota	11	99.5	10,367,776	55,458	2,000,837
Fusobacteriota	24	99.9	13,657,423	294,338	2,510,776
Mycoplasmatota	50	99.4	13,557,730	191,306	909,524
Pseudomonadota	1,850	99.7	10,723,119	167,240	4,029,573
Spirochaetota	23	99.4	10,411,464	262,078	1,861,419
Synergistota	1	98.3	1,377,705	264,125	1,852,980
Thermotogota	2	100.0	8,356,447	302,581	1,846,578
Verrucomicrobiota	4	97.4	3,363,106	113,995	2,444,956

Table 2: Archaea

Phylum	# Genome Assemblies	Avg. % Completeness	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Candidatus Thermoplasmatota	1	98.0	11,798,788	107,899	1,584,736
Euryarchaeota	36	99.3	10,538,070	175,011	2,235,608
Nitrososphaerota	1	100.0	16,828,066	640,924	1,774,962
Thermoproteota	3	99.8	8,222,310	87,974	2,743,977

Table 3: Fungi

Phylum	# Genome Assemblies	Avg. % Completeness*	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Ascomycota	208	93.4	51,686,171	563,926	2,462,588
Basidiomycota	25	90.7	35,352,995	406,137	1,624,515
Mucoromycota	1	90.2	96,084,638	838,297	398,736

Table 4: Viruses

Phylum	# Genome Assemblies	Avg. % Completeness*	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Negarnaviricota	97	92.1	1,475,745	n.a.	5,702
Peploviricota	3	63.2	8,709,399	n.a.	83,333
Uroviricota	1	83.6	721,327	n.a.	499,946
Nucleocytoviricota	1	68.8	1,436,370	n.a.	174,329
Preplasmiviricota	39	89.4	2,652,313	n.a.	34,326
Duplornaviricota	13	96.5	1,025,127	n.a.	2,400
Kitrinoviricota	12	98.2	1,741,737	n.a.	10,890
Pisuviricota	71	93.8	1,521,677	n.a.	12,765
Artverviricota	4	98.7	4,695,231	n.a.	8,216
Cossaviricota	6	81.1	1,034,691	n.a.	5,296

Tables 1-4: Summary statistics for total number of *de novo* assemblies by kingdom and phylum. Avg. % completeness is calculated by CheckM (bacteria, archaea), BUSCO (fungi), and NCBI's Viral-Genomes Database (viruses).

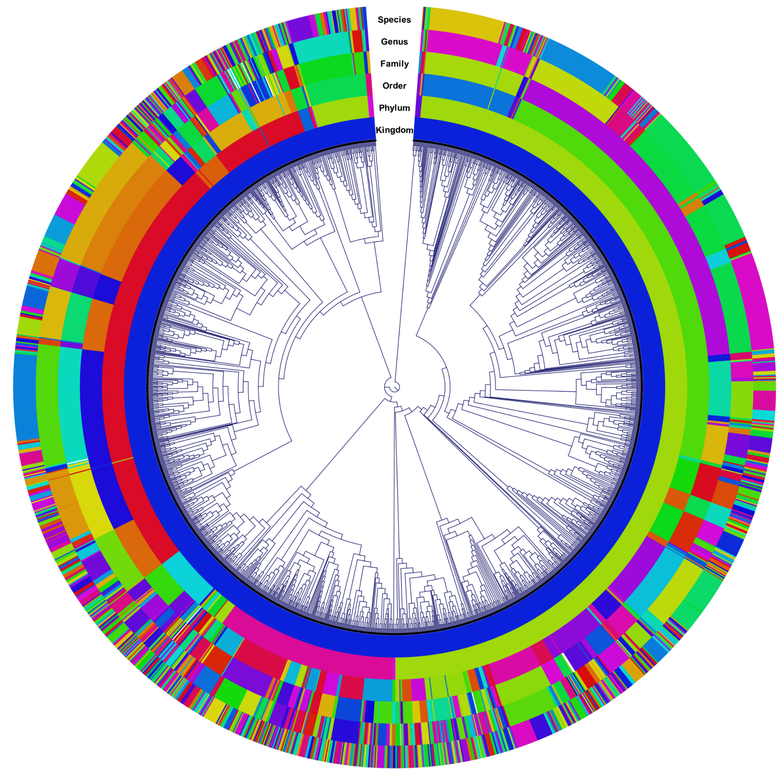


Figure 5: Kmer spectra phylogeny of all bacteria included in the ATCC Genome Portal. Colors represent progressively lower taxonomic levels.

Conclusions

- The ATCC Genome Portal provides ultra-high quality, authenticated reference genomes for ATCC bacteria, viruses, fungi, and protists.
- The database is updated monthly with new genome assemblies, annotations, and curated metadata. All data is available via the web or via our REST-API for research-use applications.

Contact

SCAN ME

Contact: Jonathan Jacobs, PhD
jjacobs@atcc.org

Scan the QR code to learn about **ATCC's Enhanced Authentication Initiative.**

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