#3960 12Jun2022

The ATCC Genome Portal

An Updated Resource for Authenticated Microbial Reference Genomes

Nikhita Puthuveetil, David Yarmosh, Patrick Ford Combs, Amy L. Reese, Corina Tabron, James Duncan, Samuel Greenfield, Robert Marlow, Stephen King, Marco Riojas, Ana Fernandes, John Bagnoli, Briana Benton, Jonathan L. Jacobs*

ATCC®

Credible leads to Incredible®

Abstract

The availability and reliability of microbial genome assemblies is essential for most microbiological research. The tension between genomic data reliability and its traceability to source materials, however, is a growing area of concern that has significant real-world impacts on diagnostics, drug discovery, and microbial ecology. While databases such as NCBI's RefSeq database have leveraged the scalability of crowd sourcing for growth, they have done so at the expense of including some unreliable, incomplete, or incorrect genomes – which have persisted in the database for decades. While the introduction of consequential data provenance policies may potentially mitigate these issues, public databases largely do not have these requirements. This creates substantial risk in the trustworthiness of individual genome assemblies, and in aggregate, for several important genomic database resources. **The ATCC Genome Portal** was created to reestablish the trustworthiness, reliability, and accuracy of genome assemblies associated with ATCC microbial strains. Currently, it includes high-quality **ATCC Standard Reference Genomes** (ASRGs) produced in-house by ATCC directly from materials sourced directly from ATCC's biorepository. **Our collection of authenticated genome references currently includes assemblies for 2,163 microbes – including 1,840 bacteria strains, 134 fungi, and 189 viruses.** 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.

Each bacterial and fungal genome is sequenced on both Illumina and Oxford Nanopore platforms, the results of which are used to produce hybrid de novo assemblies for each strain. To date, viral genomes have been produced using only Illumina sequencing. Each ASRG on the ATCC Genome Portal includes both sequencing and assembly quality metrics, annotation metrics, and metadata for each strain, such as antibiotic susceptibility, geographical origins, and phenotypic data. In addition to continually adding new genomes to the ATCC Genome Portal, we also aim include additional types of data for our strains in the future.

Here, we describe our laboratory and bioinformatics methods, the diversity of the current contents of the **ATCC Genome Portal**, it's current capabilities and new features for use by the research community. As we continue to carry out whole genome sequencing of ATCC's microbial collection, the number of reference genomes in the ATCC Genome Portal will continue to grow every month for the foreseeable future, and we encourage the research community to contact us with suggestions on taxa to prioritize in our pipeline. The ATCC Genome Portal and the data contained therein is freely available for research-use and is accessible via the web (**https://genomes.atcc.org**) or via a new REST-API.

Methods

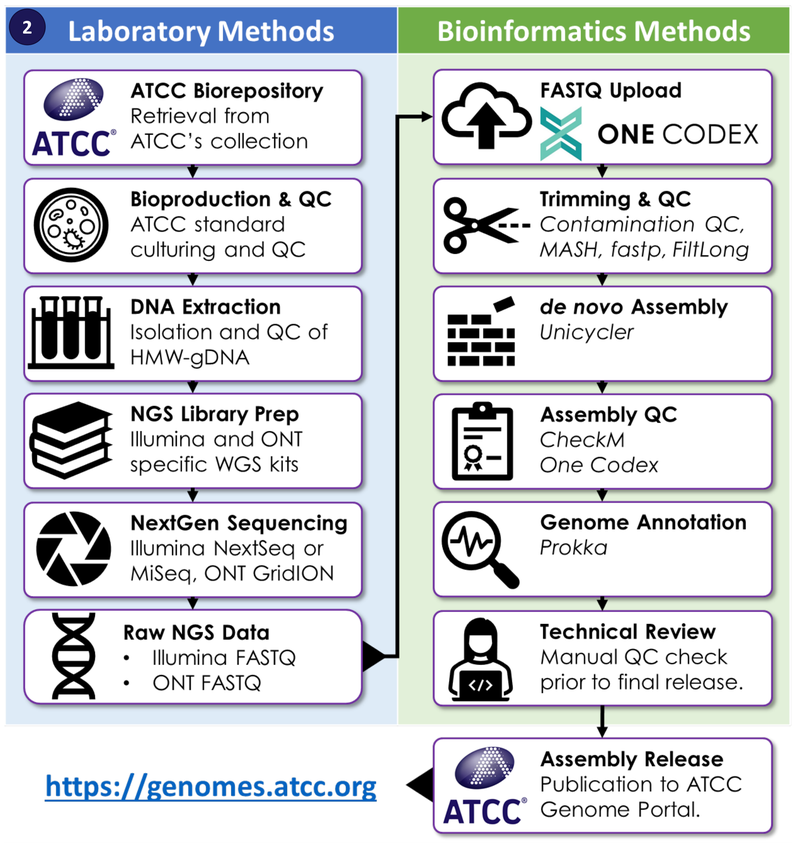

FIGURE 2: Pipeline for end-to-end genomic data provenance. Source materials were obtained directly from the ATCC biorepository and tracked through to the final assembly and genome annotation. Upfront culture conditions varied depending on the species cultured, but downstream process steps were performed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

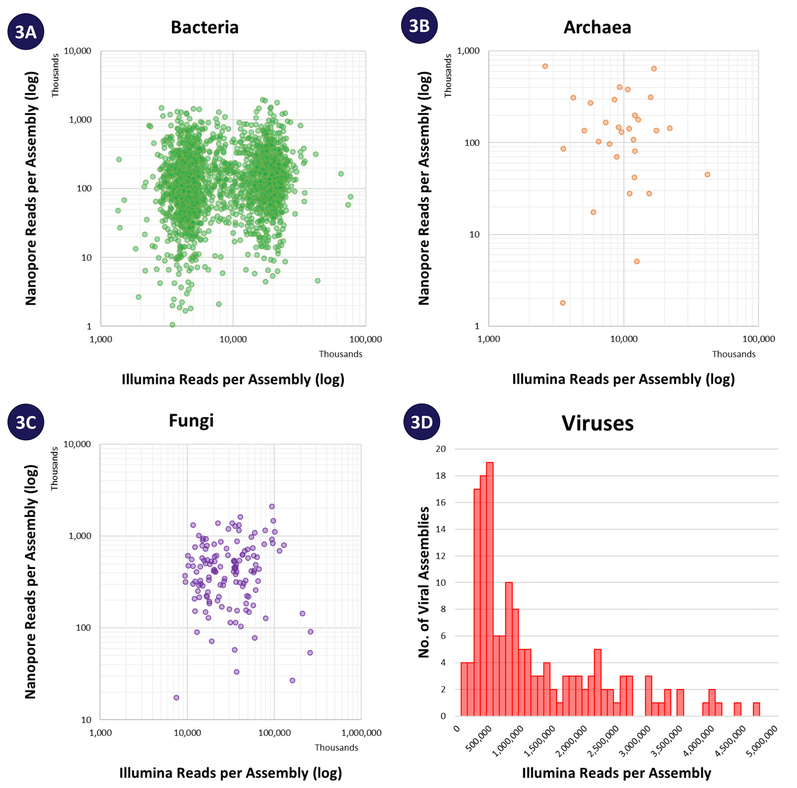

Figure 3: Total sequencing depth of both Illumina and Nanopore sequencing for all four kingdoms of organisms. A) Bacteria; B) Archaea; C) Fungi, D) Viruses. Nanopore sequencing is only performed for cellular organisms, hence the lack of this data for viral genome assemblies (2D). Additional details available at the ATCC Genome Portal under "Documentation".

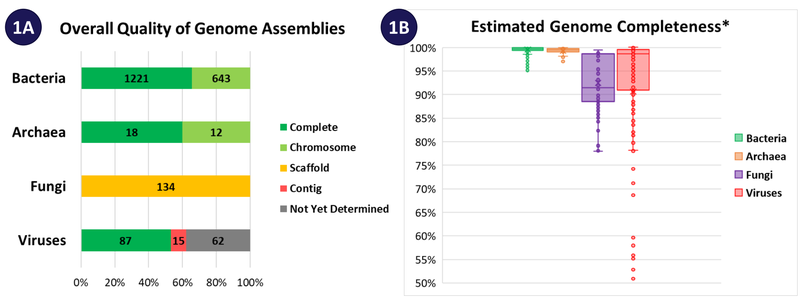

FIGURE 1: A) Overall assembly qualities for ASRGs, by kingdom, according to NCBI's assembly quality descriptors. ***B)*** Box plot of estimated genome completeness for ASRG assemblies, by kingdom. Some viral assemblies below 50% complete not shown.**In light of the high depth of sequencing and the use of both long- and short-read technologies, the lower estimations of "completeness" for fungal and viral genome assemblies are likely due to potentially significant biodiversity gaps in existing public databases.*

Results

Table 1 **BACTERIA**

Phylum	# Genome Assemblies	Avg. % Completeness	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Actinobacteria	189	99.48	11,617,673.2	176,052.8	3,105,032.6
Bacteroidetes	60	99.52	10,811,981.7	181,646.0	3,449,835.0
Chloroflexi	2	99.14	5,498,762.5	19,268.5	3,646,753.0
Cyanobacteria	6	99.71	5,499,402.7	90,236.2	4,250,611.3
Deinococcus-Thermus	8	99.41	10,641,487.3	65,023.3	2,092,967.8
Firmicutes	547	99.44	11,339,406.4	204,281.6	2,889,179.5
Fusobacteria	12	99.91	14,305,747.4	353,715.6	2,455,424.1
Proteobacteria	998	99.70	10,709,137.6	193,881.5	4,099,590.3
Spirochaetes	10	99.04	11,465,508.5	298,281.5	2,199,027.1
Synergistetes	1	98.31	1,377,705.0	264,125.0	1,852,980.0
Tenericutes	27	99.36	15,860,542.0	224,085.6	952,702.7
Thermotogae	2	100.00	8,356,447.0	302,581.0	1,846,577.5
Verrucomicrobia	2	96.94	4,664,618.5	131,834.0	2,792,539.5
Total	1,864	99.58	11,054,687.9	195,754.9	3,542,630.9

Table 2 **ARCHAEA**

Phylum	# Genome Assemblies	Avg. % Completeness	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Candidatus Thermoplasmatota	1	97.97	11,798,788.0	107,899.0	1,584,736.0
Crenarchaeota	2	99.64	8,396,301.5	83,493.5	2,602,125.5
Euryarchaeota	26	99.28	11,087,988.8	171,903.7	2,182,676.0
Thaumarchaeota	1	100.00	16,828,066.0	640,924.0	1,774,962.0
Total	30	99.28	11,123,572.2	179,510.2	2,177,117.5

Table 3 **FUNGI**

Phylum	# Genome Assemblies	Avg. % Completeness*	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Ascomycota	117	92.82	41,360,295.0	531,587.9	2,479,075.3
Basidiomycota	16	90.17	30,772,815.3	410,667.2	1,573,953.3
Mucoromycota	1	90.24	96,084,638.0	838,297.0	398,736.0
Total	134	92.48	40,504,508.9	520,083.3	2,355,476.1

Table 4 **VIRUSES**

Phylum	# Genome Assemblies	Avg. % Completeness*	Avg. Total Illumina Reads	Avg. Total Nanopore Reads	Avg. N50 per Assembly
Artverviricota	1	98.68	4,695,231.0	n.a.	8,216.0
Cossaviricota	6	78.34	942,079.2	n.a.	5,359.0
Duplornaviricota	7	97.87	722,470.9	n.a.	2,440.1
Kitrinoviricota	6	97.89	954,795.3	n.a.	11,062.8
Negarnaviricota	67	92.47	1,328,461.2	n.a.	6,514.7
Nucleocytoviricota	1	68.84	1,436,370.0	n.a.	174,329.0
Peploviricota	2	62.57	1,652,066.5	n.a.	120,948.0
Pisuviricota	46	92.55	1,266,998.3	n.a.	11,726.3
Preplasmiviricota	27	87.94	3,093,315.3	n.a.	33,739.7
Uroviricota	1	83.55	721,327.0	n.a.	499,946.0
Total	164	91.21	1,569,536.2		17,846.8

TABLES 1 – 4 (above): 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 (for viral assemblies). Nanopore sequencing is currently only performed for cellular organisms, hence the lack of this data for viral genome assemblies. Additional details available at the ATCC Genome Portal under "Documentation".

Figure 4: Reference- and alignment-free kmer phylogenetic tree for all Fungi **(A)**, Archaea **(B)**, and Viruses **(C)** currently in the ATCC Genome Portal. Colors represent progressively lower taxonomic levels, based on NCBI TaxID assignments. Complete legend hidden for space reasons, but available on request.

Reference free phylogeny highlights potential significant discrepancies in existing taxonomic classifications.

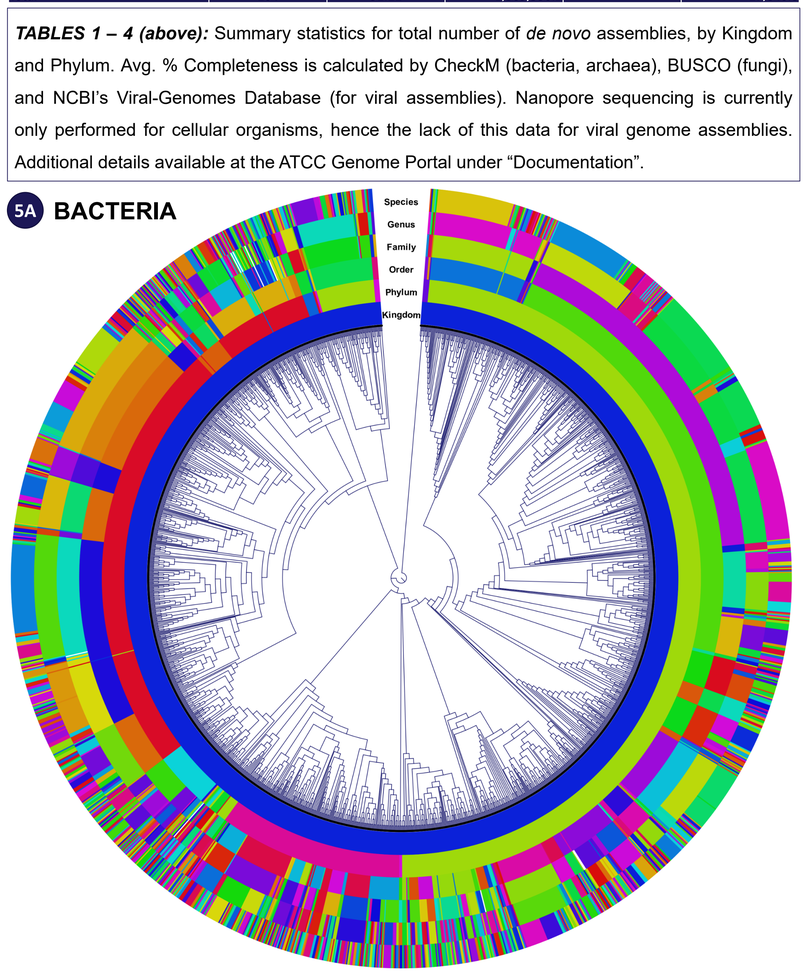

Figure 5: Kmer phylogenies of all bacteria (A), archaea (B), and virus assemblies in the ATCC Genome Portal. Colors represent progressively lower taxonomic levels (see Fig. 4).

CONCLUSIONS

- ATCC is producing ultra-high quality reference genomes for all microbial species in our collection, and providing Research-Use Only access to these data via the ATCC Genome Portal.
- Comparative genomics to public domain assemblies highlights potential widespread discrepancies in genome assembly accuracy, data provenance and traceability (see References), and potential significant taxonomic misassignments (Fig. 4).

References

1. Benton, B.; King, S.; Greenfield, S. R.; Puthuveetil, N.; Reese, A. L.; Duncan, J.; Marlow, R.; Tabron, C.; Pierola, A. E.; Yarmosh, D. A.; Combs, P. F.; Riojas, M. A.; Bagnoli, J.; Jacobs, J. L. **The ATCC Genome Portal: Microbial Genome Reference Standards with Data Provenance**. *Microbiol Resour Announc* **2021**, *10* (47), e00818-21. https://doi.org/10.1128/MRA.00818-21.

2. Yarmosh, D. A.; Lopera, J. G.; Puthuveetil, N. P.; Combs, P. F.; Reese, A. L.; Tabron, C.; Pierola, A. E.; Duncan, J.; Greenfield, S. R.; Marlow, R.; King, S.; Riojas, M. A.; Bagnoli, J.; Benton, B.; Jacobs, J. L. **Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies**. *mSphere* **2022**, e00077-22. https://doi.org/10.1128/msphere.00077-22.

CONTACT

SCAN ME

Contact: Jonathan Jacobs, PhD
jjacobs@atcc.org

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

Follow us on Twitter!

@officialATCC

@ATCCgenomics

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