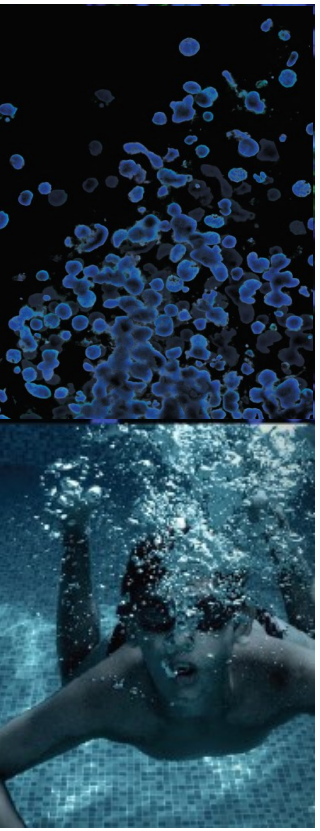
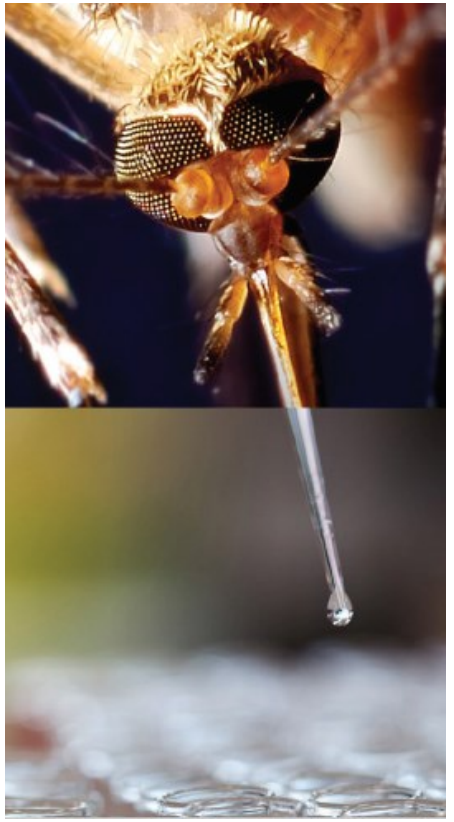
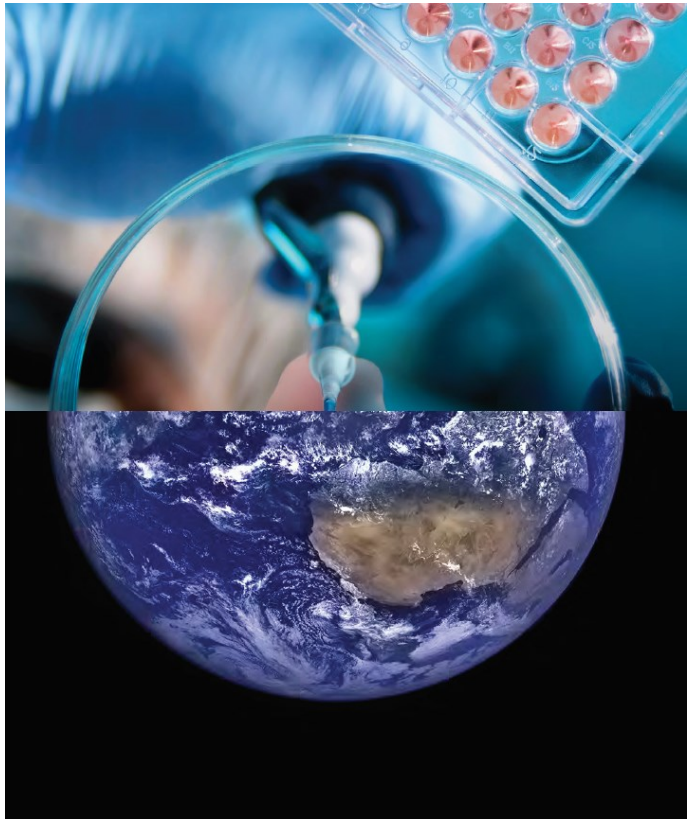
Towards Genome Authentication: The ATCC Genome Portal

John Bagnoli
August 2022





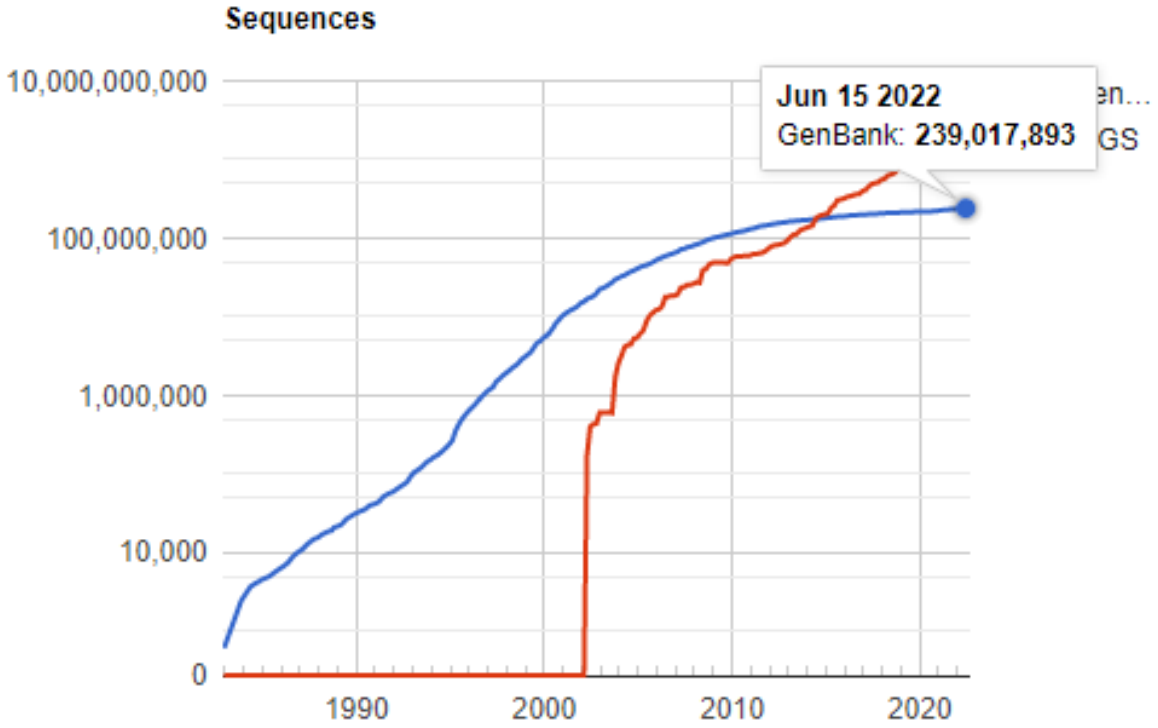


Credible Leads to Incredible®

Biological Data

The good, the bad and the ugly data

- The availability and reliability of microbial genome assemblies is essential for most microbiological research.
- Public databases have leveraged the scalability of crowd sourcing for growth, but that quantity has come at a cost.
- Consequential data provenance policies, data curation experts and more safeguards may potentially mitigate these issues.

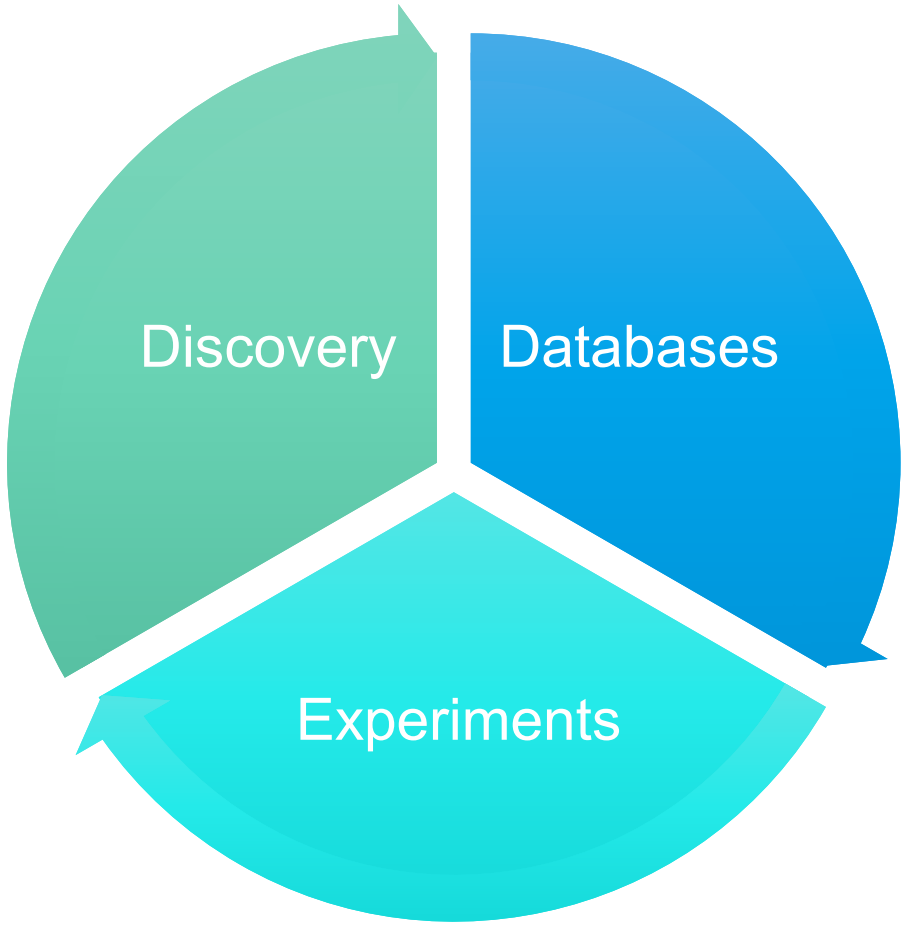


Snapshot (https://www.ncbi.nlm.nih.gov/genbank/statistics/) June 2022

Database size is doubling in growth approximately every 18 months

Data Impact

Why it Matters



The tension between genomic data reliability and its traceability to source materials is a growing area of concern that has significant real-world impacts across multiple research areas.

Yet 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.

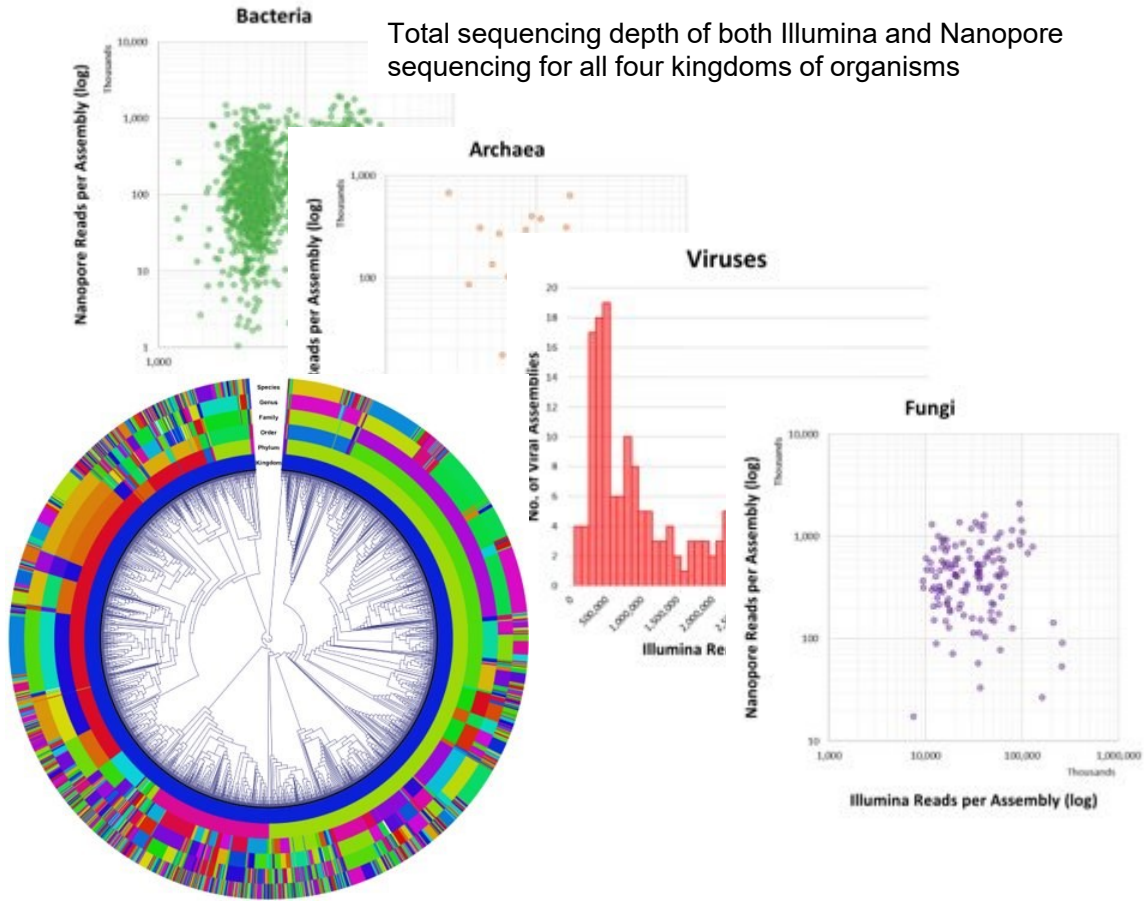
The need for well characterized quality genomics data is crucial for life science research. Laboratories often leverage publicly available data as a cornerstone for their experimental design.

Authentication Initiative

Standards for our Standards

- American Type Culture Collection (ATCC) started an internal Authentication Initiative to address this problem
- ATCC created the Genome Portal to establish trustworthiness, reliability, and accuracy of genome assemblies associated with ATCC materials.
- Currently, it includes high-quality ATCC Standard Reference Genomes (ASRGs) produced in-house by ATCC directly from materials sourced from ATCC's biorepository.

Total sequencing depth of both Illumina and Nanopore sequencing for all four kingdoms of organisms

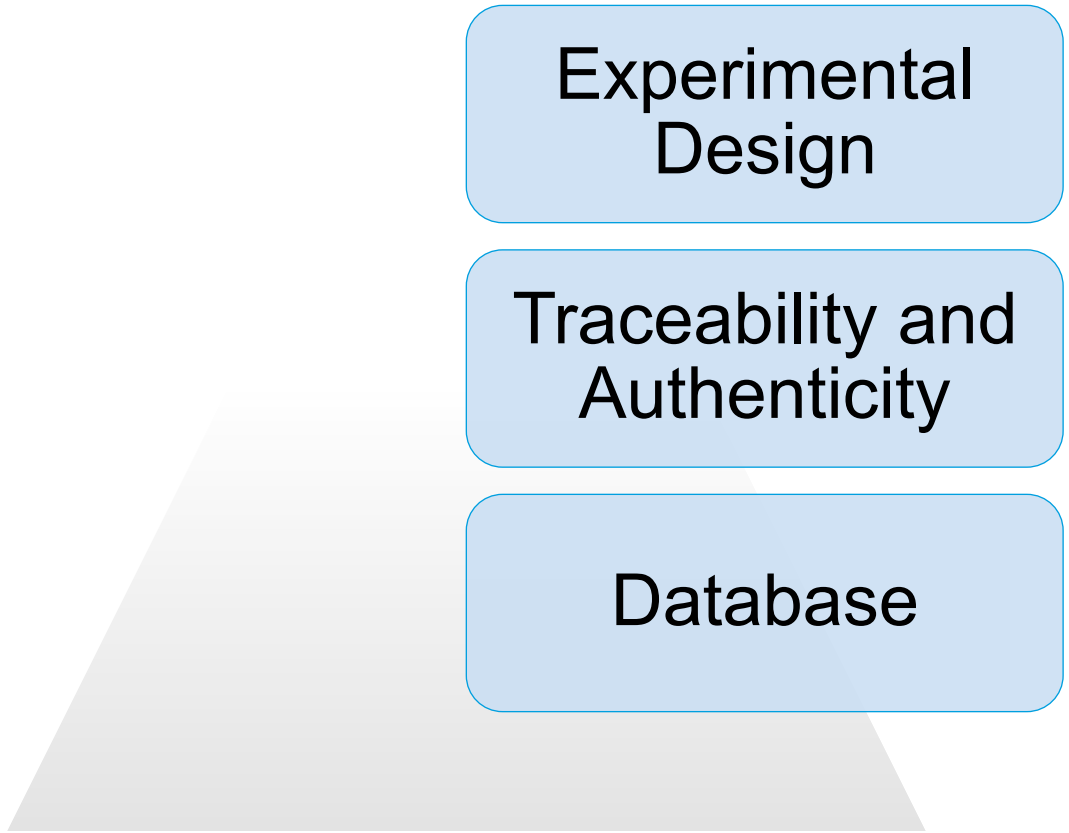


Kmer phylogenies of all bacteria in the ATCC Genome Portal

Database Standards

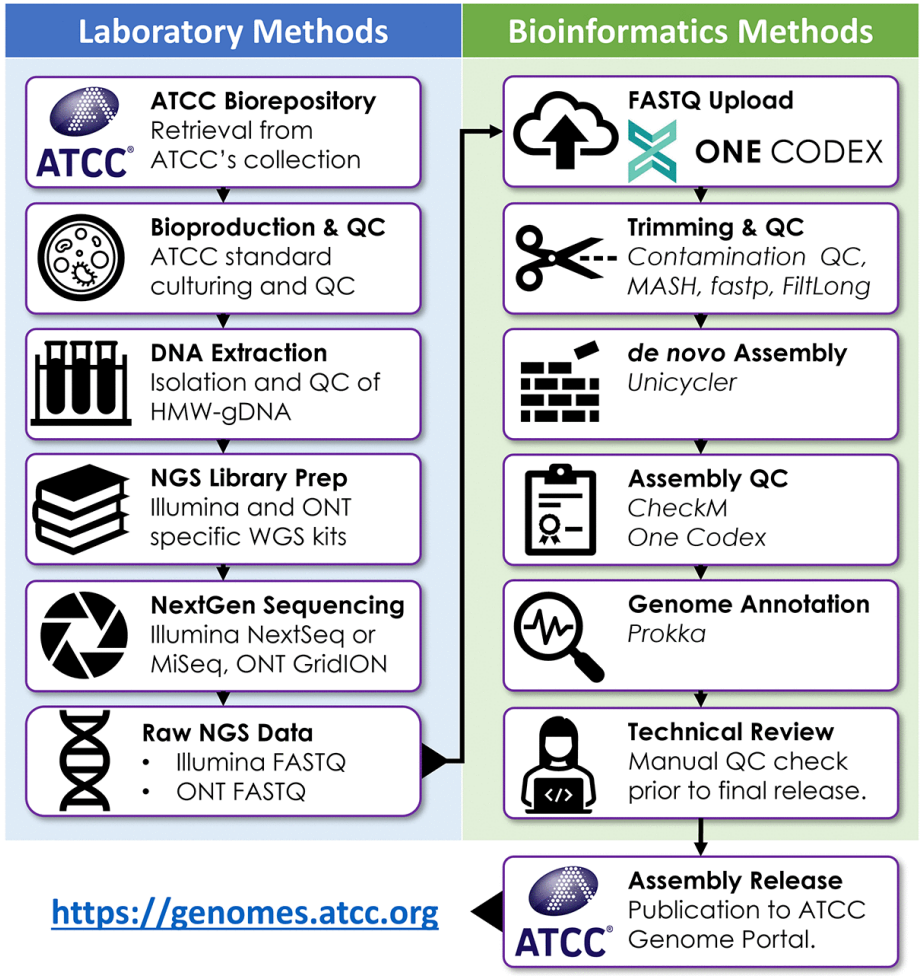
The Foundations of Good Data

- There is very high interest for us in helping establish standards
- Databases underlie many studies, including those involving the microbiome



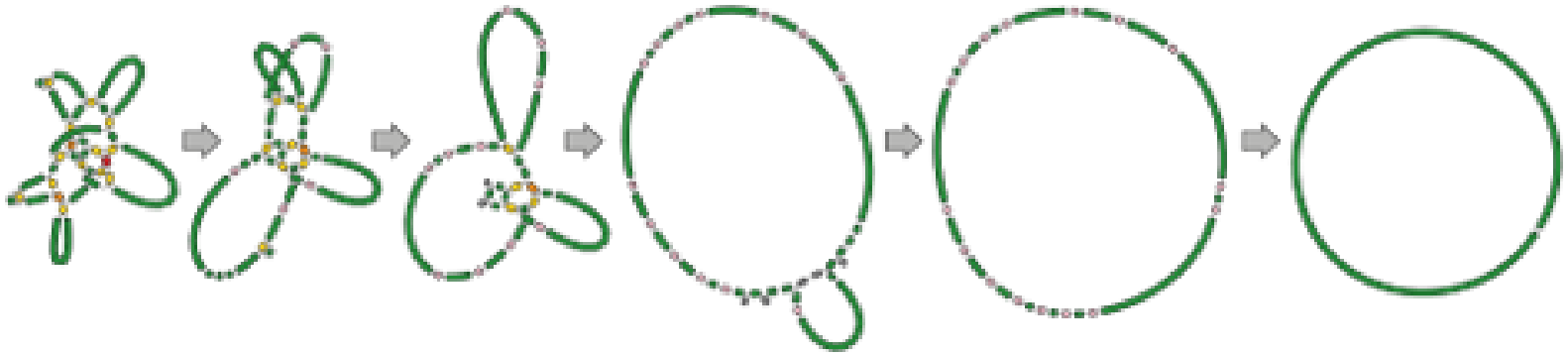
Data Generation

The Sequencing Process in Brief



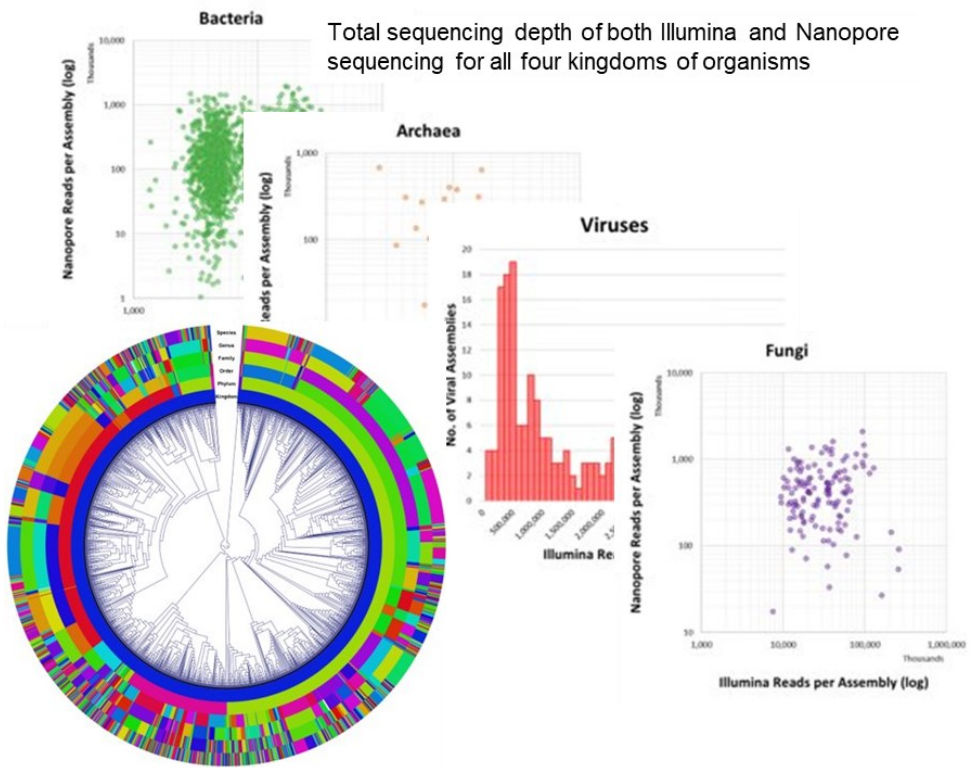
- Combination of laboratory techniques and bioinformatic tools.
- Hybrid – Illumina plus Oxford Nanopore sequencing technologies
  - Long read for good assemblies
  - Short reads to ensure quality
- No single person can be an expert across the Tree of Life
  - Manual review combined with automated quality check

Why use hybrid data?



- Genome Assembly is a jigsaw puzzle
- Illumina reads can have trouble resolving edges and nodes (where contigs overlap) which results in a “tangled” assembly
- ONT long-reads are used to resolve the connections between contigs with more contextual information for the algorithm
  - Repeats, inverted sections and a lot more

The Genome Portal



Total sequencing depth of both Illumina and Nanopore sequencing for all four kingdoms of organisms

Kmer phylogenies of all bacteria in the ATCC Genome Portal

- The content of the ATCC Genome Portal is updated every month with new genome assemblies
- Materials are traceable, including entire mock communities
- Characterization of ATCC standards – you know what you get in the tube
- Looking for community suggestions – you can request ATCC sequencing something in the collection

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.

JOIN THE CHALLENGE

We invite researchers in academic and industry laboratories to submit proposals of your innovative research using ATCC Next-generation Sequencing (NGS) Standards for a chance to win free vials. This research can span any field of interest—human or environmental. The submission period will take place between August 15 – September 23, 2022, so start submitting your proposals!

Submit your proposal at www.atcc.org/InnovationChallenge

Thank you

Questions?