AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

DATABASES AND SOFTWARE

The ATCC Genome Portal: Microbial Genome Reference Standards with Data Provenance

Briana Benton,[a] Stephen King,[a] Samuel R. Greenfield,[a] Nikhita Puthuveetil,[a] Amy L. Reese,[a] James Duncan,[a] Robert Marlow,[a] Corina Tabron,[a] Amanda E. Pierola,[a] David A. Yarmosh, Jr.,[a] Patrick Ford Combs,[a] Marco A. Riojas,[a] John Bagnoli,[a] Jonathan L. Jacobs[a]

[a]American Type Culture Collection, Manassas, Virginia, USA

ABSTRACT Lack of data provenance negatively impacts scientific reproducibility and the reliability of genomic data. The ATCC Genome Portal (https://genomes.atcc.org) addresses this by providing data provenance information for microbial whole-genome assemblies originating from authenticated biological materials. To date, we have sequenced 1,579 complete genomes, including 466 type strains and 1,156 novel genomes.

The data provenance (origins) of genome assemblies found in International Nucleotide Sequence Database Collaboration (INSDC) databases increasingly have gaps, errors, and omissions, which collectively create real-world challenges for interpretation and scientific reproducibility relying on genomic data (1–7). Although depositor assurances and INSDC quality pipelines create some trust regarding these data, these approaches fall short of providing complete data provenance information and authentication, since there is no guarantee that the physical biomaterials are available or that the associated metadata are accurate. Further complications include gaps in the recorded chain of custody of the source materials, undisclosed phenotypic differences, variability in naming conventions, and lack of standardized data formats, all of which impact the data provenance and the reliability of public genomic data.

To address the issue of provenance and attribution for genome references, the American Type Culture Collection (ATCC) is systematically sequencing, assembling, and annotating its entire microbial collection. The goal is to provide the research community with provenance information and authentication between the biological source materials and reference genome assemblies derived from them. To date (18 October 2021), 1,579 genome assemblies are available for research use via the ATCC Genome Portal (AGP) (https://genomes.atcc.org), a publicly accessible web portal and genome database. This total includes 1,396 bacterial, 74 fungal, and 109 viral genome assemblies, of which 73% (1,156 genome assemblies) represent genome assemblies of novel organisms and strains.

All genome assemblies in the AGP are derived from authenticated materials available from the ATCC and are part of the ATCC Enhanced Authentication Initiative. In general, this includes one or more of the following: assessment of colony morphology, Gram staining, metabolic profiling, antibiotic susceptibility testing, biochemical reactivity testing, 16S rRNA sequencing, matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS), and whole-genome next-generation sequencing.

Bacteria and fungi included in the AGP (1,475 genome assemblies [~93%]) are sequenced on both Illumina and Oxford Nanopore Technologies (ONT) sequencing platforms, whereas viruses are sequenced only on Illumina systems. All reads are quality filtered, trimmed, and down-sampled to an estimated 100× maximum coverage using MASH (8). ONT reads are filtered and trimmed using Filtlong, whereas fastp is used for Illumina reads (9, 10). Viral and bacterial genomes are assembled *de novo* using SPAdes, and fungal genomes are assembled using MaSuRCA with Flye (11–13). Postassembly quality control and annotation

Editor J. Cameron Thrash, University of Southern California

Address correspondence to Jonathan L. Jacobs, jjacobs@atcc.org.

Received 16 August 2021
Accepted 25 October 2021
Published 24 November 2021

of each genome are performed using CheckM (bacteria), Prokka (bacteria), BUSCO (fungi), and scripts from One Codex (viruses) (14–17).

The AGP represents one of the first microbial genome databases to provide genomic data provenance information between the source materials and their genome assemblies. Our goal is to continue our practice of updating the AGP monthly with new assemblies. In addition, in the near future we will begin to include phenotypic, cell imaging, and lot-specific bioproduction data, as well as additional omic data as they become available.

Data availability. The details of specific laboratory methods, DNA extraction and sequencing quality thresholds, bioinformatic pipelines, software parameters, the REST application programming interface (API), bulk downloads of raw data, and the databases used are available online (https://github.com/ATCC-Bioinformatics/AGP-Resource-Announcement).

ACKNOWLEDGMENT

The development of the AGP is financially supported solely by the ATCC.

REFERENCES

1. Rajesh A, Chang Y, Abedalthagafi MS, Wong-Beringer A, Love MI, Mangul S. 2021. Improving the completeness of public metadata accompanying omics studies. Genome Biol 22:106. https://doi.org/10.1186/s13059-021-02332-z.
2. Pettengill JB, Beal J, Balkey M, Allard M, Rand H, Timme R. 2021. Interpretative labor and the bane of non-standardized metadata in public health surveillance and food safety. Clin Infect Dis 73:1537–1539. https://doi.org/10.1093/cid/ciab615.
3. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L, Sadzewicz L, Nadendla S, Klimke W, Hatcher E, Shumway M, Aldea DL, Allen J, Koehler J, Slezak T, Lovell S, Schoepp R, Scherf U. 2019. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nat Commun 10:3313. https://doi.org/10.1038/s41467-019-11306-6.
4. Bagheri H, Severin AJ, Rajan H. 2020. Detecting and correcting misclassified sequences in the large-scale public databases. Bioinformatics 36:4699–4705. https://doi.org/10.1093/bioinformatics/btaa586.
5. Leipzig J, Nüst D, Hoyt CT, Soiland-Reyes S, Ram K, Greenberg J. 2021. The role of metadata in reproducible computational research. Patterns 2:100322. https://doi.org/10.1016/j.patter.2021.100322.
6. Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P. 2021. The critical importance of vouchers in genomics. Elife 10:e68264. https://doi.org/10.7554/eLife.68264.
7. Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. 2018. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 46:D48–D51. https://doi.org/10.1093/nar/gkx1097.
8. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.
9. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
10. Wick R, Menzel P. 2019. Filtlong: quality filtering tool for long reads. https://github.com/rrwick/Filtlong.
11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
12. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669–2677. https://doi.org/10.1093/bioinformatics/btt476.
13. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
14. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
15. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
16. Minot SS, Krumm N, Greenfield NB. 2015. One Codex: a sensitive and accurate data platform for genomic microbial identification. bioRxiv 027607. https://doi.org/10.1101/027607.
17. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.