ATCC® | Credible leads to Incredible™

Technical Document

Genome Portal

As life science research progresses, the quality of data becomes increasingly more important. With our Enhanced Authentication Initiative, we aim to enrich the characterization of our biological collections and provide you with the whole-genome sequences of the specific, authenticated materials you need to generate credible data.

The purpose of this technical documentation is to outline the features of the ATCC Genome Portal as well as provide comprehensive descriptions of the DNA extraction, sequencing, and bioinformatic methods we use to produce high-quality, reference-grade genomes.

ATCC Genome Portal

The ATCC Genome Portal offers more than just a collection of reference-grade bacterial genomes originating from authenticated ATCC materials. It is a platform where users can interactively browse genomic data and metadata that is both searchable and indexed.

Portal Features

- Browse and download whole-genome sequences and annotations of ATCC microbial products
- Search for nucleotide sequences or genes within published genomes
- Search for genomes by taxonomic name, taxonomic level, isolation source, ATCC catalog number, type strain status, or biosafety level
- View genome assembly statistics and quality metrics
- Identify the relatedness of published genomes by total genome alignment
- Purchase the corresponding authenticated ATCC source material

Our Approach to Bacterial Genome Sequencing

After decades of bacterial DNA sequencing, a plethora of techniques exist to sequence and assemble bacterial genomes.[1,2] At ATCC, we are setting the scientific standard in best practices for bacterial whole-genome sequencing as part of our Enhanced Authentication Initiative.

Recent innovation in third-generation sequencing[3,4] have now made it possible to produce complete reference-grade bacterial genomes by combining highly accurate Illumina® short reads with the revolutionary scaffolding ability of Oxford Nanopore Technologies® (ONT) ultra-long reads via so-called hybrid assembly techniques[5,6] (for additional details see section: Genome Assembly).

The ATCC bacterial whole-genome sequencing workflow is an optimized methodology designed to achieve complete, circularized (when biologically appropriate) bacterial genomic elements by using the hybrid assembly technique. This methodology comprises five primary steps:

1. **Extraction** of DNA from authenticated ATCC strains
2. **Sequencing** of this DNA
3. **Assembly** of sequencing data into a genome
4. **Annotation** of the resultant genome
5. **Estimation of relatedness** between a genome and all other genomes in our collection

Each step is accompanied by rigorous quality control methods and criteria to ensure that the data proceeding to the next step are the highest quality possible. Only the data that pass all quality control criteria are published to the ATCC genome portal. While ATCC materials undergo extensive quality control while being grown, a description of these processes is outside the scope of this document. For more information, see our whitepaper on ATCC prokaryotic authentication.

In the sections below, the methods and/or bioinformatic tools used to accomplish each step are described alongside relevant scientific citations supporting that approach. In addition, methods and/or bioinformatic tools used to measure quality control criteria are described alongside relevant scientific citations supporting the use of that measurement.

DNA Extraction

High-quality DNA extraction is the critical starting point to creating a complete reference-grade genome. ATCC uses several proprietary protocols to obtain high-molecular-weight extractions from our microbial portfolio; the method chosen is dependent on the organism undergoing extraction. For select ATCC microbes, ATCC is now making next-generation sequencing (NGS)–ready DNA available for customer purchase.

Whole-Genome Sequencing

To generate the best quality sequencing data for our genome assemblies, we perform a single DNA extraction and sequence the DNA on both Illumina and ONT sequencing platforms.

Illumina Sequencing

Illumina libraries are prepared using the latest library preparation kits available. Libraries are subsequently sequenced on an Illumina instrument, producing a paired-end read set per sample. The degree of sample multiplexing is based on the estimated genome size of a given organism and the amount of data necessary to generate at least 100X coverage of the genome with the Illumina read set. Resultant reads are adapter trimmed using the adapter trimming option on the Illumina instrument. Periodic updates to the instruments' software are performed when they are made available by the manufacturer to ensure that the latest version of instrument software is used for basecalling and adapter trimming for a given sequencing date.

Oxford Nanopore Technologies Sequencing

ONT libraries are prepared using the latest DNA sequencing kits available, then sequenced on an ONT instrument with the latest flow cell version available. The degree of sample multiplexing is based on the estimated genome size of a given organism. Flow cells are run on the instrument for at least 12 hours. Periodic updates to the instruments' software are performed when they are made available by the manufacturer to ensure that the latest version of ONT software is used for sequencing and basecalling for a given sequencing date.

After basecalling, all resultant FASTQs are combined and then demultiplexed using either porechop or qcat, with barcode removal settings turned on.

Illumina Data Quality Control

Illumina read sets commonly contain flanking low-quality regions and portions of Illumina adapter sequence; removing these regions can substantially improve genome assemblies.[7] To accomplish this, we perform a round of adapter removal and quality filtering. This also ensures removal of adapter sequences otherwise missed by Illumina software.

After Illumina read sets undergo quality and adapter trimming, we assess the quality of the read set by using FastQC. Illumina reads must pass the following quality control:

1. Median Q score, all bases > 30
2. Median Q score, per base > 25
3. Ambiguous content (% N bases) < 5%

Oxford Nanopore Technologies Data Quality Control

ONT ultra-long reads are critical for scaffolding over the low-complexity regions of bacterial genomes during hybrid assembly, but they have limited influence in determining base identity given enough Illumina coverage.[6,8] Given the lower quality of ONT sequencing data, all data was trimmed and filtered for low quality regions. The quality control metrics used across all ONT read sets produced are:

1. Minimum mean Q score, per read > 10
2. Minimum read length > 5000

To perform this quality control step, we employ NanoFilt on demultiplexed ONT read sets, in addition to barcode sequence removal during demultiplexing.

Read-Based Contamination Quality Control with One Codex

ATCC employs state-of-the-art methods to detect contamination during the growth phase of our product production. To compliment this approach, we use the One Codex microbial genomics platform[10] to perform read-level *k*-mer–based taxonomic classification and estimation of strain abundances on our processed Illumina read sets, which represent a highly-accurate snap shot of a given DNA extraction. A minimum of 1,000,000 Illumina reads per sequenced sample is required to undergo such analysis; Illumina read sets otherwise passing quality control criteria but possess less than 1,000,000 reads are sent for re-sequencing. When an Illumina read set is confirmed as an isolate by the One Codex platform, all read sets from that extraction continue to genome assembly. Please note that the results of this reads-based analysis are not currently presented on the portal but that all published genomes have passed our stringent thresholds for purity

Genome Assembly

Hybrid Assembly

Hybrid assembly is a state-of-the-art technique that uses both highly accurate Illumina short reads and ultra-long scaffolding ONT reads. In general, this technique begins with an optimized Illumina assembly. The longest of these resultant contigs are then assembled alongside the ONT reads; this combined assembly then undergoes multiple rounds of both long-read and short-read polishing. Because occasional sequencing and assembly artifacts appear as small contigs in the final assembly (so-called "chaff" contigs[11]), non-contiguous contigs less than 1000 bp with low relative coverage are then removed to produce the final assembly.

Genome Assembly Quality Control

Genetic Element Contiguity

Most bacterial genome assemblies available in public databases are of "draft" quality,[12] and thus single genetic elements (*e.g.*, chromosomes, plasmids, and other types of mobile genetic elements) are split between multiple contigs with little to no structural arrangement data. To be published to the ATCC genome portal, single genetic elements in ATCC reference-grade genomes must be assembled into a single contig. When the assembly process supports circularization of a contig (as in the case of most bacterial chromosomes and some mobile genetic elements), they are reported as such. Genome assemblies that possess multiple contigs that the genome assembly process recognizes are contiguous—but have unresolved structural relationships—are currently excluded from the genome portal.

Illumina Read Set Coverage

Although the depth of Illumina reads required is influenced by numerous factors (including, but not limited to, bacterial strain),[13,14] Illumina read sets should be sufficient to cover the entire genome to obtain the most accurate base determination.[6] To account for variance in distribution of coverage per base, we require a minimum of 100X average coverage for Illumina reads.

CheckM Completeness and Contamination

To ensure our assembly process has correctly captured the entirety of a given strain's genome, and to confirm the absence of contamination from the assembly, we pass finalized assemblies through CheckM.[15] Briefly, CheckM uses a set of Hidden Markov Models (HMMs) from phylogenetically close reference genomes to determine if the query assembly contains all expected HMMs as predicted by the reference genomes (a percentage called "CheckM completeness"), and what percent of the query's HMMs differ in copy number or come from reference genomes that are phylogenetically distant (called "CheckM contamination"). We required final assemblies to have completeness values ≥ 95% and contamination values ≤ 5% (*e.g.*, within the margin of error for 0% completeness and contamination, which indicates them as "excellence reference sequences" according to the authors of CheckM).

Genome Annotation

There are currently several approaches for bacterial genome annotations.[16,17,18] As such, we make our finalized genome assembly FASTA files available for download from our genome portal and encourage our customers to conduct their own custom annotations of the ATCC reference-grade genomes if they so choose. However, we also recognize the need for a rapidly accessible annotation in a common format for those looking to perform immediate data analysis at the gene level. To address these needs, we provide a default genome annotation for ATCC reference-grade genomes with prokka.[17] Briefly, prokka relies on a number of tools to annotate CDS, rRNA, tRNA, signal leader peptides, and non-coding RNA. For CDSs, prokka leverages the UniProt,[19] RefSeq,[20] Pfam,[21] and TIGRFAM[22] databases to assign protein identity. On the genome portal, all annotated CDSs include their EC number and UniProt ID as reported by prokka.

Estimation of Genome Relatedness

ATCC's reference-grade bacterial genomes have even greater analytical power when considered in context of other closely related genomes in our database. To measure relatedness between our published genomes, we implement the most widely used approach: average nucleotide identity (ANI). In this framework, ANI values greater than 95% between two genomes indicates that these genomes are derived from members of the same prokaryotic species.[23]

Interactive Genome Search

A *k*-mer based nucleotide search is used to power the interactive genome search feature on the portal. The sequence search matches all *k*-mers (*k*=31) in the query against all available ATCC reference genomes and highlights portions of the sequence that match. The minimum requirement is matching 40 *k*-mers and 80% of the sequence to call a hit. Search results are listed in descending order by percent of matching *k*-mers.

References

1. Niedringhaus TP, *et al*. Landscape of next-generation sequencing technologies. *Analytical Chemistry*, 83(12): 4327–4341, 2011. PubMed: 21612267
2. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*, 13(12): 787–794, 2015. PubMed: 26548914
3. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5): 518–524, 2016. PubMed: 27153285
4. Jain M, *et al*. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1): 239, 2016. PubMed: 27887629
5. Maio N, *et al*. The REHAB Consortium. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *BioRxiv*, 530824, 2019.
6. Wick RR, *et al*. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, 13(6): e1005595, 2017. PubMed: 28594827
7. Del Fabbro C, *et al*. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE*, 8(12): e85024, 2013. PubMed: 24376861
8. Shomorony I, Courtade T, Tse D. Do read errors matter for genome assembly? *2015 IEEE International Symposium on Information Theory* (ISIT), 919–923, 2015.
9. Koren S, *et al*. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, 14(9): R101, 2013. PubMed: 24034426
10. Minot SS, Krumm N, Greenfield NB. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *BioRxiv*, 27607, 2015.
11. Salzberg SL., *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3), 557–567, 2012. PubMed: 22147368
12. Land M, *et al*. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2): 141–161, 2015. PubMed: 25722247
13. Desai A, *et al*. Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. *PLoS ONE*, 8(4): e60204, 2013. PubMed: 23593174
14. Pightling AW, Petronella N, Pagotto F. Choice of Reference Sequence and Assembler for Alignment of *Listeria monocytogenes* Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. *PLoS ONE*, 9(8): e104579, 2014. PubMed: 25144537
15. Parks DH, *et al*. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7): 1043–1055, 2015. PubMed: 25977477
16. Overbeek R, *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1): D206–D214, 2014. PubMed: 24293654
17. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14): 2068–2069, 2014. PubMed: 24642063
18. Zhao Y, *et al*. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3): 416–418, 2012. PubMed: 22130594
19. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1): D204–D212, 2015. PubMed: 25348405
20. Tatusova T, *et al*. RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research*, 42(D1): 3872, 2014. PubMed: 25824943
21. Finn RD, *et al*. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1): D222–D230, 2014. PubMed: 24288371
22. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1), 371–373, 2003. PubMed: 12520025
23. Goris J, *et al*. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1): 81–91, 2007. PubMed: 17220447

These products are for laboratory use only. Not for human or diagnostic use. ATCC products may not be resold, modified for resale, used to provide commercial services or to manufacture commercial products without prior ATCC written approval.