

TECHNICAL DOCUMENTS

GENOME PORTAL

As life science research progresses, the quality of data becomes increasingly important. As part of our initiative to enhance the authentication of our products, we aim to enrich the characterization of our biological collections by providing the whole-genome sequences of the specific, authenticated materials you need to generate credible data.

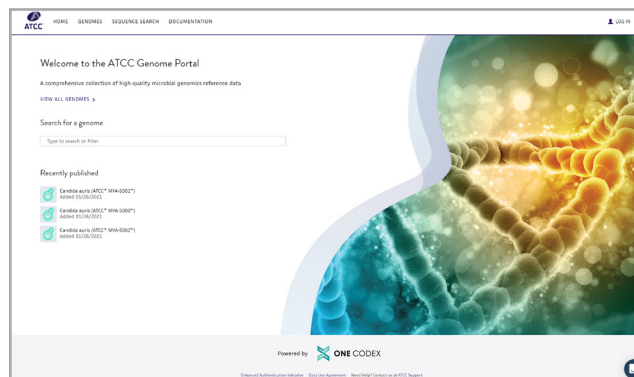
The purpose of this technical documentation is to outline the features of the ATCC Genome Portal and provide comprehensive descriptions of the DNA or RNA extraction, sequencing, and bioinformatic methods we use to produce high-quality, reference-grade genomes.

ATCC GENOME PORTAL

The ATCC Genome Portal offers more than just a collection of reference-grade bacterial, viral, fungal, or protist genomes originating from authenticated ATCC materials. It is a platform where users can interactively browse genomic data and metadata.

PORTAL FEATURES

- Browse and download whole-genome sequences and annotations for a variety of ATCC products. The collection currently includes thousands of bacterial, fungal, viral, and protist genomes. Plus, we consistently release new assemblies every quarter.
- 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, biosafety level, or certain tags. Current tag options include Type Strain and MSA Component.
 - Type strain: Strain with official standing in prokaryotic nomenclature
 - MSA Component: Component of an ATCC NGS Standard
- View genome assembly statistics and quality metrics.
- Identify the relatedness of published genomes by total genome alignment.
- Purchase the corresponding authenticated ATCC source material.
- Access the portal from the command line through an API (https://github.com/ATCC-Bioinformatics/genome_portal_api)



OUR APPROACH TO GENOME SEQUENCING

After multiple decades of development in nucleic acid (DNA and RNA) sequencing, a plethora of techniques exist to sequence and assemble microbial genomes.¹⁻⁴ At ATCC, we are setting the scientific standard in best practices for whole-genome sequencing.

Recent innovations in second- and third-generation sequencing⁵⁻⁷ have now made it possible to produce complete reference-grade microbial genomes and improve the assembly contiguity of large and highly heterozygous fungal 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^{8,9} (for additional details see sections: Whole-Genome Sequencing and Genome Assembly).

The ATCC microbial whole-genome sequencing workflow is an optimized methodology designed to achieve complete, circularized (when biologically appropriate), and contiguous genomic elements by using short-read (RNA virology collection) and hybrid (bacteriology, mycology, and DNA virology collections) assembly techniques. This methodology comprises five primary steps:

- 1 **Extraction** of nucleic acids from authenticated ATCC strains
- 2 **Sequencing** of the nucleic acids
- 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 the product sheet for each product.

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.

NUCLEIC ACID EXTRACTION

High-quality DNA or RNA 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.



WHOLE-GENOME SEQUENCING

To generate the best quality sequencing data for our genome assemblies, we perform a single DNA or RNA extraction. ATCC uses both Illumina and Oxford Nanopore Technology (ONT) platforms for sequencing. See collection sections below for greater detail.

ILLUMINA SEQUENCING

Illumina (DNA and RNA) libraries are prepared using the latest and most reliable library preparation kits available. Libraries are subsequently sequenced on an Illumina instrument (MiSeq® or NextSeq2000®), 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 depth 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 base-calling and adapter trimming for a given sequencing date.

OXFORD NANOPORE TECHNOLOGIES SEQUENCING

ONT libraries are prepared using the latest and most reliable DNA sequencing kits available, then sequenced on an ONT instrument (GridION) with the latest and most reliable 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 48 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 base-calling for a given sequencing date.

After base-calling, all resultant FASTQs are combined and then demultiplexed using [MinKNOW](#) with the 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.¹⁰ To accomplish this, we perform a second round of adapter and quality filtering using [fastp](#). This also ensures the 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 and fungal genomes during hybrid assembly, but they have limited influence in determining base identity given enough Illumina depth.⁷⁻⁹ 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 > 1000 bp. To perform this quality control step, we employ [Filtlong](#) 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 complement this approach, we use the [One Codex microbial genomics platform](#)¹¹ to perform read-level *k*-mer-based¹² taxonomic classification and estimation of strain abundances on our processed Illumina read sets, which represent a highly-accurate snapshot 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 possessing fewer 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.¹³ For bacterial assemblies, this technique uses [Unicycler](#), which begins with an optimized Illumina assembly. The longest of these Illumina-based contigs are then assembled alongside the ONT reads; this combined assembly then undergoes multiple rounds of both long-read and short-read polishing.⁸ Because occasional sequencing and assembly artifacts appear as small contigs in the final assembly (so-called “chaff” contigs¹⁴), non-contiguous contigs less than 1000 bp with low relative depth are then removed to produce the final assembly. Please note that prior to bacterial genome assembly, the genome size is estimated using [MASH](#), and the high-quality Illumina and ONT reads are down-sampled to 150X and 30X depth respectively based on estimated genome size.

For fungal assemblies, we down-sample the reads as for the bacterial assemblies, and then use the MaSuRCA (hybrid assembly algorithm combines Illumina and ONT reads to construct long and accurate mega-reads) pipeline with the Flye assembler.¹⁵ [MaSuRCA](#) was chosen for its strengths with large genomes.⁹

GENOME ASSEMBLY QUALITY CONTROL

Illumina Read Set Coverage

Although the depth of Illumina reads required is influenced by numerous factors (including, but not limited to, microbial strains),¹⁶⁻¹⁷ Illumina read sets should be sufficient to cover the entire genome to obtain the most accurate base determination.¹⁸ To account for variance in distribution of coverage per base, we require a minimum of 100X average depth for Illumina reads.

Bacterial Completeness and Contamination

To ensure our bacterial 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](#).¹⁹ Briefly, CheckM uses a set of Hidden Markov Models (HMMs) from phylogenetically close bacterial and archaeal reference genomes to determine if the query assembly contains all expected HMMs as predicted by the reference genomes (a percentage called “CheckM completeness”), and it evaluates 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 $\geq 95\%$ and contamination values $\leq 5\%$ (eg, within the margin of error for 0% completeness and contamination, which indicates them as excellent reference sequences according to the authors of CheckM).

Mycology Completeness and Contamination

For mycology genomes, we estimate completeness using [BUSCO](#).²⁰ BUSCO is a tool/database combo widely used in the mycology field that examines the presence of a selection of universal single-copy orthologs for quantitative completeness calculations. We use [fungi-specific databases](#) where orthologs must be identified in at least 90% of the fungal species, and no single copy ortholog can be entirely missing from any sub-clade in the databases. Unlike CheckM, BUSCO does not calculate % contamination. We require fungal assemblies to have a completeness value of $\geq 80\%$.

Viral Genome Assembly and Quality Assessment

As viruses are co-cultured with their host, viral DNA or RNA sequencing data may contain reads from both the host and the virus, and de novo assemblies may contain contigs from both the host and viral genome.²¹ In order to produce an assembly containing contigs of a single virus; host reads can be removed or contigs can be binned taxonomically.²² Taxonomic binning is performed by aligning reads or contigs to One Codex’s curated NCBI Reference Sequence Database. Reads or contigs that align to “cellular organisms” are binned as nonviral, while those that do not are binned as viral. For our approach, high-quality viral reads are used for de novo assembly using [SPAdes](#).²³ To achieve the goal of obtaining complete assemblies for a single virus, the contig binning approach was used. Contigs that align to the *Escherichia coli* bacteriophage Phi-X 174 genome are excluded as this is used as a DNA spike-in for Illumina sequencing.

In addition to the problem of taxonomic binning, viral genomes are diverse in structure with many viruses having multipartite genome segments; the genome of Influenza A virus, for example, consists of 8 separate strands of RNA.²⁴ To determine whether an assembly contains all the necessary segments, a curated database of complete viral genomes and segment information was constructed. After taxonomic binning, contigs are then aligned to the [Viral Genomes-NCBI-NIH](#) database to apply segment labels, segment depth, and percent identity to the closest reference.

GENOME ANNOTATION

BACTERIOLOGY GENOME ANNOTATION

There are currently several approaches for bacterial genome annotations.²⁵⁻²⁷ 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 NCBI's Prokaryotic Genome Annotation Pipeline (PGAP).²⁶ While bacterial assemblies from 2019 to 2023 were initially annotated using Prokka,²⁵ new and existing genomes are currently being annotated using NCBI's PGAP. PGAP combines ab initio gene prediction algorithms with homology-based methods. PGAP leverages the Protein Family Models collection for structural and functional annotation. This collection is composed of Hidden Markov Model (HMM), Blast (BlastRules), and Conserved Domain Database-based architectures (CDDs) to assign names, gene symbols, publications, and EC number to the proteins that meet criteria for protein family inclusions. On the ATCC Genome Portal, all annotated CDSs include their EC number and UniProt ID as reported by PGAP.

MYCOLOGY GENOME ANNOTATION

During completeness calculations for mycology genomes, BUSCO²⁰ generates annotations of universal single-copy orthologs, which we make available in the genome portal. BUSCO uses Augustus (trained on BUSCO databases), tBLASTn, and HMMER3 to automatically predict and annotate single-copy coding regions of mycological genomes according to their closest relatives on fungi-specific databases.

VIRAL GENOME ANNOTATION AND VARIANT DETECTION

Viral assemblies draw gene annotations from the closest reference sequence in Viral Genomes-NCBI-NIH databases by using a customized python script.

To call genomic variants, the depth-masked, SPAdes de novo assembly is aligned to a reference sequences using MAFFT,²⁹ which is a tool for multiple sequence alignments. Briefly, multiple sequence alignments are converted to a table of variants by using custom scripts. The table of variants is then joined to the reference assembly's genome annotation to produce a table of variants and their overlapping annotations. Variants that extend to the end of a segment sequence are excluded as they are likely truncations and not true biological variants. Variants that contain entirely ambiguous nucleotides in the reference or alternate sequence are also excluded from reporting.

CALCULATION OF ASSEMBLY LEVEL

Assemblies are ranked into different categories based on the NCBI Assembly Level definitions:

Bacteriology

- Complete – Hybrid assemblies with $\geq 95\%$ completeness as estimated by CheckM. All contigs are fully circularized.
- Scaffold – Hybrid assemblies with $\geq 95\%$ completeness as estimated by CheckM.
- Contig – All remaining assemblies. Contig-level assemblies are not published in the genome portal.

Mycology

- Complete – Hybrid assemblies with $\geq 80\%$ completeness as estimated by BUSCO, with a single contig per chromosome.
- Scaffold – Hybrid assemblies with $\geq 80\%$ completeness as estimated by BUSCO.
- Contig – All remaining assemblies. Contig-level assemblies are not published in the genome portal.

Virology

- Complete – Assemblies with $\geq 80\%$ completeness with all segments present and each represented by a single contig. Only complete assemblies are published in the genome portal.
- Scaffold – Not used.
- Contig – All remaining assemblies. Contig-level assemblies are not published in the genome portal.

ESTIMATION OF GENOME RELATEDNESS

ATCC's reference-grade microbial 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).³⁰ In this framework, ANI values greater than 95% between two genomes indicate that these genomes are derived from members of the same bacterial or archaeal species.³¹ Additionally, related members of the genus are determined by NCBI taxonomy.


INTERACTIVE GENOME SEARCH


A k -mer based nucleotide search is used to power the interactive genome search feature on the portal.³² 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 Shendure J, et al. DNA sequencing at 40: past, present and future. *Nature* 550(7676): 345-353, 2017.
- 2 Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* 13(12): 787-794, 2015.
- 3 Sharma KK. Fungal genome sequencing: basic biology to biotechnology. *Crit Rev Biotechnol* 36(4): 743-759, 2016.
- 4 Ladner JT, et al. Standards for sequencing viral genomes in the era of high-throughput sequencing. *mBio* 5(3): e01360-14, 2014.
- 5 Giani AM, et al. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* 18: 9-19, 2020.
- 6 Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol* 34(5): 518-24, 2016.
- 7 Amarasinghe SL, et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21(1): 30, 2020.
- 8 Wick RR, et al. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol*. 13(6): e1005595, 2017.
- 9 Zimin AV, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* 27(5): 787-792, 2017.
- 10 Del Fabbro C., et al. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* 8(12): e85024, 2013.
- 11 Minot SS, Krumm N, Greenfield NB. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *bioRxiv*, 2015.
- 12 Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 20(4): 1125-1136, 2019.
- 13 Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics* 21(1): 631, 2020.
- 14 Salzberg SL, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* 22(3): 557-567, 2012.
- 15 Zimin AV, et al. The MaSuRCA genome assembler. *Bioinformatics* 29(21): 2669-2677, 2013.
- 16 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.
- 17 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.
- 18 Sims D, et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 15(2): 121-132, 2014.
- 19 Parks DH, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25(7): 1043-1055, 2015.
- 20 Seppy M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* 1962: 227-245, 2019.
- 21 Miller JR, et al. A host subtraction database for virus discovery in human cell line sequencing data. *F1000Res* 7: 98, 2018.
- 22 Daly GM, et al. Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing yData. *PLoS One* 10(6): e0129059, 2015.
- 23 Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5): 455-477, 2012.
- 24 Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of thousands of viral genomes. *Elife* 7, 2018.
- 25 Overbeek R, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42(Database issue): D206-14, 2014.
- 26 Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14): 2068-2069, 2014.
- 27 Zhao Y, et al. PGAP: pan-genomes analysis pipeline. *Bioinformatics* 28(3): 416-418, 2012.
- 28 Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* 20(4): 1160-1166, 2019.
- 29 Palmer J, Stajich J. Funannotate v1.8.15: Eukaryotic genome annotation (v1.8.15), 2020. Zenodo.<https://doi.org/10.5281/zenodo.4054262>; <https://github.com/nextgenusfs/funannotate>
- 30 Jain C, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9(1): 5114, 2018.
- 31 Goris J, et al. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57(Pt 1): 81-91, 2007.
- 32 Marchet C, et al. Data structures based on *k*-mers for querying large collections of sequencing data sets. *Genome Res* 31(1): 1-12, 2021.

 10801 University Boulevard
Manassas, Virginia 20110-2209

 703.365.2700

 703.365.2701

 sales@atcc.org

 www.atcc.org

GP-032024-v08

©2024 American Type Culture Collection. The ATCC trademark and trade name, and any other trademarks listed in this publication are trademarks owned by the American Type Culture Collection unless indicated otherwise. Illumina, Nextera, NextSeq, and MiSeq are trademarks or registered trademarks of Illumina, Inc. Oxford Nanopore Technologies, EPI2ME, GridION, and MinKNOW are registered trademarks of Oxford Nanopore Technologies Limited.

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.