ATCC

TECHNICAL DOCUMENTS

# THE ATCC GENOME PORTAL: OUR APPROACH TO MICROBIAL WHOLE-GENOME SEQUENCING

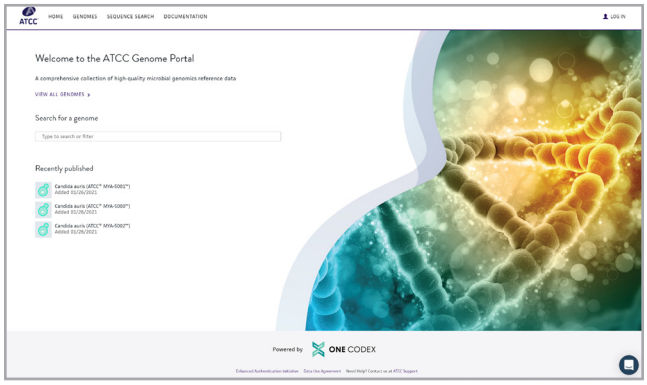
As life science research progresses, the quality of data becomes increasingly important. As part of our initiative to enhance the authentication of our products, we aim to enrich the characterization of our biological collections by providing whole-genome sequences of the specific, authenticated materials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

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 microbial genomes. Beyond microbial reference genomes, the portal also offers whole-exome and RNA-seq data from our mammalian cell lines. For additional information on cell line data, please refer to our technical document: The ATCC Genome Portal: Our Approach to Cell Line Whole-Exome and RNA Sequencing.

## 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, strain alias, 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)

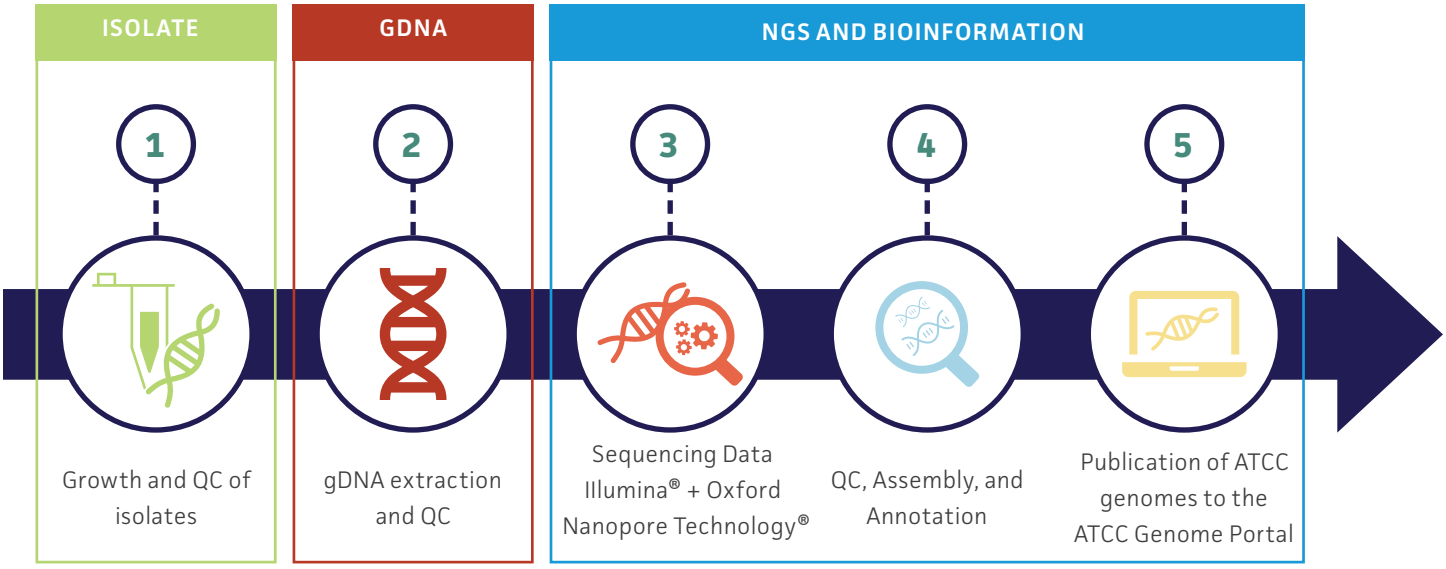
## OUR APPROACH TO GENOME SEQUENCING

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 sequencing[5-7] 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) *de novo* 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 (currently only bacteriology and mycology collections)
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.

The following sections describe the processes and bioinformatic tools applied at each step of our methodology, alongside quality control criteria and supporting scientific citations. 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. Instrument software is periodically updated when new versions are released by the manufacturer to ensure the latest software is used for base-calling and adapter trimming on the sequencing date.

### OXFORD NANOPORE TECHNOLOGIES SEQUENCING

ONT libraries are prepared using the latest and most reliable DNA sequencing kits available, and are 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–based[12] 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; samples with 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 while the results of our reads-based analysis are not currently presented on the portal, all published genomes have passed this stringent threshold for purity.

# GENOME ASSEMBLY

## HYBRID *DE NOVO* ASSEMBLY

In contrast to reference-based assemblies, which depend on an existing genome for alignment, *de novo* assemblies reconstruct genomes directly from raw reads. This approach reduces mapping bias, enables detection of novel sequences and structural variants, and identifies potential contaminants.[13] Hybrid *de novo* assembly is a state-of-the-art technique we employ that uses both highly accurate Illumina short reads and ultra-long scaffolding ONT reads.[13] For constructing bacterial genomes, 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 and the combined assembly undergoes multiple rounds of both long-read and short-read polishing.[8] For cases where additional curation may be needed to generate a reference-quality assembly, Flye may also be used as an approach to generate an ONT-first assembly, which is then polished with Illumina data. The assembler used is logged and maintained in the JSON metadata behind a genome on the portal. Because occasional sequencing and assembly artifacts appear as small contigs in the final assembly (so-called "chaff" contigs[14]), non-contiguous contigs less than 1000 bp with low relative depth are 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 a minimum of 100X and 40X, respectively, based on estimated genome size and *k*-mer frequency.

For fungal assemblies, the reads are down-sampled as for the bacterial assemblies. MaSuRCA—a hybrid assembly algorithm that combines Illumina and ONT reads to construct long and accurate mega-reads—is utilized with the cabog and/or 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, the specific microbial strain),[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 within the assembly, we evaluate finalized assemblies through CheckM.[19] Briefly, CheckM uses a set of Hidden Markov Models (HMMs) from phylogenetically related 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"); it then evaluates what percent of the query's HMMs differ in copy number or originate from reference genomes that are phylogenetically distant (i.e., "CheckM contamination"). We require final assemblies to have completeness values ≥ 95% and contamination values ≤ 5% (eg, within the margin of error for 0% completeness and contamination), which indicates "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 ≥ 80%.

### Virology Completeness an Contamination

To collect metrics regarding viral completeness, CheckV[21] is used to compare against NCBI GenBank viral genomes and calculate relative length, which is relatively invariant amongst genera.[22] Alongside the detection of inverted terminal repeats, direct terminal repeats, and integrated provirus sequences, CheckV calculates and reports a completeness value used to assess genome completeness. Well studied viral genomes must have a completeness score of ≥ 80% to be considered for publication.

To assess contamination, all assembled contigs in our virology pipeline are queried against the NCBI nt database using blastn. All contigs that classify as off-target virus that may be co-assembled with our target genome are manually separated and investigated in further detail. No virology genomes are published to the ATCC Genome Portal with evidence of off-target contamination.

### 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 unfiltered *de novo* assemblies may contain contigs from both the host and viral genome.[23] In order to produce an assembly containing contigs of a single virus, host reads can be removed or contigs can be binned taxonomically.[24] We first utilize a strategy of removing host

reads by retaining reads that do not map to a database of genomic sequences comprising all possible cell-line backgrounds used at ATCC. For our approach to RNA viruses, high-quality Illumina-sequenced, viral cDNA reads are used for *de novo* assembly with SPAdes.[25] Similar to how we assemble contiguous bacterial and fungal genomes, for our dsDNA and ssDNA virology items, we utilize a *de novo* hybrid approach. For viruses, Flye may also be used to generate an ONT-first scaffolding that is polished with the Illumina data. To achieve the goal of obtaining complete assemblies for a single virus, a final contig classification and extraction approach is also used. Contigs are queried against the NCBI nt database with blastn to identify and remove potential noise in the assembly. This methodology also allows us to remove residual host contigs that may have been co-assembled. Contigs that additionally align to the *Escherichia coli* bacteriophage Phi-X 174 genome are excluded as standard QC. PhiX174 is a known Illumina sequencing spike-in and has been frequently reported (e.g., >1,000 genomes in GenBank; ~10% published in literature: https://pmc.ncbi.nlm.nih.gov/articles/PMC4511556/). 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.[26] 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.[27-29] 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. 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).[29] Notably, while our bacterial assemblies from 2019 to 2023 were innitally annotated using Prokka,[28] our new and existing genomes are annotated using NCBI's PGAP, which 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. For more information on annotation of the resistome, see the resistome section below.

### MYCOLOGY GENOME ANNOTATION

To provide comprehensive feature annotations for genomes in our mycology collection, we utilize Funannotate to generate standardized GenBank annotation files for publication on the ATCC Genome Portal. Funannotate integrates multiple evidence-based and ab initio approaches to predict and annotate gene structures, ensuring high-quality functional characterization. The pipeline leverages training of gene models using Augustus,[32] GeneMark,[33] SNAP,[34] and GlimmerHMM,[35] while incorporating homology searches through Diamond[36] and HMMER3[37] for protein domain identification. Functional annotation is enriched using eggNOG[38] orthology assignments and BUSCO[20] assessments for completeness. This multi-tool strategy enables accurate gene prediction, functional categorization, and quality control, providing researchers with additional genomic resources to be utilized in downstream comparative and functional analyses.

### CALCULATION OF ASSEMBLY LEVEL

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

**Bacteriology**

- Complete - Hybrid assemblies with ≥ 95% completeness as estimated by CheckM. All contigs are fully circularized.
- Scaffold – Hybrid assemblies with ≥ 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 ≥ 80% completeness as estimated by BUSCO, with a single contig per chromosome.
- Scaffold – Hybrid assemblies with ≥ 80% completeness as estimated by BUSCO.
- Contig – All remaining assemblies. Contig-level assemblies are not published in the genome portal.

**Virology**

- Complete – Assemblies with ≥ 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).[39] 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.[40] 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.[41] 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.

## METHYLATION

All methylation BED files on the ATCC Genome Portal are generated using ONT's Dorado basecaller. However, prior to bed file creation, there are several pre-processing steps that are performed on the data based on the ONT flow cell that was used to generate the ONT sequencing data.

### DATA CONVERSION TO POD5 FILE FORMAT

The first pre-processing step is a data conversion step to ensure the input files are in a POD5 format. The data conversion methods slightly vary depending on 1) the version of the ONT flow cell and 2) the year that the ATCC sample was sequenced.

**ONT Data Generated Using a R9.4.1 Flow Cell**

For ONT data generated using a R9.4.1 flow cell, the raw sequencing data is output as a fast5 file. For samples that were sequenced in 2018 or 2019, the data will be stored as single-read fast5 files. Using a set of custom Bash scripts, the data will be converted to multiread fast5 files, converted to a POD5 format, re-basecalled, and de-multiplexed to generate final de-multiplexed POD5 files.

For samples that were sequenced after 2019 on the R9.4.1 flow cell, the data will be stored as multiread fast5 files. These files will be converted to a POD5 format using the ONT POD5 conda package (https://github.com/nanoporetech/pod5-file-format).

**ONT Data Generated Using a R10.4.1 Flow Cell**

For ONT data generated using a R10.4.1 flow cell, raw sequencing data is output as a POD5 file. For samples that were sequenced in 2024, these POD5 files are sufficient to move directly to methylation calling and will bypass this pre-processing step. For samples from 2025 and onward, the POD5 data will be multiplexed. During the data conversion step, the POD5 file will be re-basecalled and de-multiplexed using a set of custom Bash scripts.

### SELECTION OF METHYLATION BASECALLING MODELS

The final pre-processing step is the selection of the Dorado methylation basecalling models according to each flow cell type. For R9.4.1 flow cells, v3.3 HAC methylation model is used: 5mCG_5hmCG. For R10.4.1 flow cells, the v5.0.0 HAC basecalling models are used: 4mC_5mC, 5mCG_5hmCG, 5mC_5hmC, and 6mA. During methylation calling, the 5mC_5hmC and 6mA models will be run together.

### METHYLATION CALLING USING DORADO

Once the pre-processing steps are completed, methylation calling can begin. As input, the selected methylation basecalling model and each ATCC sample's processed POD5 file, AGP reference FASTA file, and barcoding kit name are provided to ONT's Dorado basecaller. The resulting bam file is then sorted, indexed, and used as input for modkit pileup to generate the final bedMethyl files. For samples generated with R9.4.1 flow cells, only 1 bedMethyl file will be created. For samples generated with R10.4.1 flow cells, up to 3 bedMethyl files will be created (one for each model).

After the BED files are generated, a README JSON file is created for each sample containing sample, genome, and methylation metadata. The BED file(s) and JSON are then packaged into a zip folder and published to the AGP.

Additional details related to methylation and a list of all AGP genomes with available methylation data can be found on the AGP_methylation Github page.

## RESISTOME ANNOTATION BY AMRFINDER, CARD AND RESFINDER

During the bacterial annotation process, an additional layer of scrutiny is created to identify antimicrobial resistance (AMR) markers using AMRFinderPlus,[42] RGI/CARD,[43] and Resfinder[44] collectively to identify acquired AMR genes, mutations and some efflux/overexpression mechanisms.

NCBI's AMRFinderPlus uses a protein homology search using HMMs (Hidden Markov Models) and BLASTP against a curated reference database of AMR proteins/genes. CARD RGI (Resistance Gene Identifier) utilizes BLAST and HMMs against the CARD reference database (protein homologs, mutations, and rRNA mutation models). ResFinder uses k-mer alignment or BLASTN against a database of acquired resistance gene sequences at the nucleotide-level.

Custom scripts collate all the information together into a single output, provided as an excel file along with a JSON with relevant metadata. The results of the AMR databases are collated together in context with traditional annotations and as separate outputs. The excel sheets contain: Collated AMR hits with PGAP annotations, each individual software's results, PGAP Annotations, all annotations merged with all AMR database output columns (under WARNING-FULL), and a PanISa analysis. PanISa[45] (pan-Insertion Sequence analyzer) provides information on detected insertion sequences by analyzing soft-clipped reads and flanking direct repeats in short-read alignments to identify and reconstruct IS elements without a reference database.

The identification of these AMR genes are also displayed on the AGP in the Genome Browser under "AMR genes" for any bacterial item where these annotations are found.

## 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 Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol 1962: 227-245, 2019.
21 Nayfach S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39: 578–585, 2021.
22 Roux S, et al. Minimum Information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37(1): 29–37, 2019).
23 Miller JR, et al. A host subtraction database for virus discovery in human cell line sequencing data. F1000Res 7: 98, 2018.
24 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.
25 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.
26 Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of thousands of viral genomes. Elife 7, 2018.
27 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.
28 Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14): 2068-2069, 2014.
29 Zhao Y, et al. PGAP: pan-genomes analysis pipeline. Bioinformatics 28(3): 416-418, 2012.
30 Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20(4): 1160-1166, 2019.
31 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
32 Stanke M, et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32: W309-312, 2004.
33 Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26(4): 1107-1115, 1998.
34 Johnson AD, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24): 2938-2939, 2008.
35 Majoros WH, et al. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16): 2878-2879, 2004.
36 Buchfink B, et al. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1): 59-60, 2015.
37 Mistry J, et al. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41(12): e121, 2013.
38 Huerta-Cepas J, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47(D1): D309-D314, 2019.
39 Jain C, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9(1): 5114, 2018.
40 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.
41 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.
42 Feldgarden M, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11(1): 12728, 2021.
43 Alcock BP, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48(D1): D517-D525, 2020.
44 Florensa AF, et al. ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genom 8(1): 000748, 2022.
45 Treepong P, et al. PanISa: An R Package for Ab Initio Detection of Insertion Sequences from Short-Read Sequencing Data. Bioinformatics 35(2): 310–312, 2018.

10801 University Boulevard
Manassas, Virginia 20110-2209

703.365.2700 703.365.2701 sales@atcc.org www.atcc.org

**GP-112025-v09**