**Query ATCC whitelist genomes**

Here, we demonstrate how to use *ATCCfinder* to download and query against ATCC reference genomes. ATCC provides a set of 9 reference genes that are available to access for free with ATCC account credentials.

The following protocol should be performed on a computer with internet access. Downloaded sequences will be ~70 megabytes uncompressed. Note that if computer storage is limited, *ATCCfinder* can handle compressed query and target files. Ability to use command line for executing code is required. A data table visualization software, such as Microsoft Excel or LibreOffice Calc, is recommended but not necessary for viewing results files.

Download query sequence from NCBI-Nucleotide database

1. Navigate to the NCBI-nucleotide search page: <https://www.ncbi.nlm.nih.gov/nuccore/>
2. Paste accessions (Table 1) into the search bar (Figure 1). Note that NCBI will automatically convert newline-delimited accessions (Table 1) into tab-delimited ones.

|  |  |  |  |
| --- | --- | --- | --- |
| **Taxonomy** | **rRNA** | **RefSeq Accession** | **Sequence Length** |
| *Bacillus subtilis* | 16s | NR\_112116.2 | 1550 |
| *Cryptococcus neoformans* | ITS | NR\_171785.1 | 554 |
| *Enterococcus faecalis* | 16s | NR\_115765.1 | 1483 |
| *Escherichia coli* | 16s | NR\_114042.1 | 1467 |
| *Lactobacillus fermentum* | 16s | NR\_104927.1 | 1502 |
| *Listeria monocytogenes* | 16s | NR\_044823.1 | 1469 |
| *Pseudomonas aeruginosa* | 16s | NR\_114471.1 | 1489 |
| *Saccharomyces cerevisiae* | ITS | NR\_111007.1 | 752 |
| *Salmonella enterica* | 16s | NR\_119108.1 | 1539 |
| *Staphylococcus aureus* | 16s | NR\_118997.2 | 1552 |

**Table 1. Amplicon sequence accessions from species present in the ZymoBIOMICS microbial community standard sample.** Columns describe genus and species classification (Taxonomy), amplicon sequence type (rRNA), NCBI-RefSeq database accession (RefSeq Accession), and the length of the amplicon sequence in base pairs (Sequence Length).



**Figure 1. Search query in NCBI-nucleotide database.** Accession codes are entered into the search bar (outlined, red).

1. Search results will proceed directly to the results page (Figure 2). Select the FASTA format option and then download the sequences by selecting “Send to” and ensuring “Complete Record” and “File” are selected (Figure 2).



**Figure 2. NCBI-Nucleotide search result page for amplicon accessions.** Select “Send to” (outlined, red) to expand the drop-down menu, select “FASTA” format (outlined, orange), and select “Create File” (outlined, yellow) to initiate the download prompt. Note that it may be necessary to select “File” under “Choose Destination” in order to view “Format” and “Create File” options.

1. The selected sequence will now be downloaded to your computer. A prompt will appear to specify download location and filename. The downloaded file will hereafter be referred to as “amplicon\_sequences.fasta”. The query sequence file has now been downloaded.

Use *ATCCfinder* to download ATCC target database sequences

ATCC provides a selection of reference genomes as publicly available for download without requiring a subscription to their genome database. Provided accession numbers, *ATCCfinder* can download specific ATCC reference genomes in the same way that it broadly can access and download the entire dataset, providing a demonstration of this capability.

|  |  |  |  |
| --- | --- | --- | --- |
| **Taxonomy** | **ATCC Accession** | **Assembly ID** | **Strain** |
| *Aspergillus brasiliensis* | 36efe8d13467435c | 4c8a0d3f62a640d0 | 16404 |
| *Candida albicans* | 0fdc61a44a8f4582 | 94ac833fd37f43c4 | 10231 |
| *Escherichia coli* | b9d91f150db449de | 5f85abaf737847cd | 8739 |
| *Escherichia coli bacteriophage MS2* | 231e1b568ee14e91 | 1a86b2e5e0a343a5 | 15597-B1 |
| *Exiguobacterium indicum* | 080e20c7507b453b | 864afe7737034735 | TSD-220 |
| *Human adenovirus 5* | 9fdf2eb2010942a7 | 58cd1ee1118e4361 | VR-5 |
| *Moraxella catarrhalis* | 6547647b5a0f460d | ae97d9f8e3b44073 | 23246 |
| *Severe acute respiratory syndrome coronavirus 2* | ec8c390b482542ee | 374e6b9622444ea1 | VR-1986 |
| *Staphylococcus aureus* | 79f43b45f79b4abc | f96252fe15ad4c2e | 6538 |

**Table 2. ATCC whitelist references.** Publicly available ATCC whitelist reference sequences, with columns describing taxonomic classification of each reference (Taxonomy), ATCC genome identifier accession code (ATCC accession), ATCC genome catalog name (Assembly ID), and strain name (Strain).

1. By default, *ATCCfinder* downloads all ATCC reference genomes. A file may be specified by the parameter “--atcc\_ids” to only download and compile specific ATCC references. Here we use this parameter to only download publicly available whitelist ATCC references. Begin by creating a text file “atcc\_whitelist\_genomes.txt”, containing ATCC Accessions (Table 2), separated by newline characters (Figure 3).



**Figure 3. ATCC accession file for only downloading specific ATCC references.** ATCC accessions are listed on each line, separated by a newline character.

1. ATCC requires an API key to access their genomes. Instructions for setting up a compatible API key may be found here: <https://docs.onecodex.com/en/articles/5812163-atcc-genome-portal-api-guide>.
2. Download *ATCCfinder*. ATCCfinder is available on GitHub at: <https://github.com/lanl/ATCCfinder> and Zenodo at: https://doi.org/10.5281/zenodo.15178103. Navigate to the web page and select the “Code” drop down menu, then “Download ZIP” to download the software (Figure 4).



**Figure 4. Download ATCCfinder software from GitHub.** Select the “Code” dropdown (outlined, red) and then “Download ZIP” (outlined, orange) to download *ATCCfinder*.

1. Decompress the downloaded ZIP folder for *ATCCfinder* and enable execution permissions:

$ unzip ATCCfinder-main.zip -d ATCCfinder-main

$ chmod -R +x ATCCfinder-main/

1. Create and activate a software environment for executing *ATCCfinder* download:

## Download genome\_portal\_api package

git clone https://github.com/ATCC-Bioinformatics/genome\_portal\_api.git

## Define environment

mamba create -n ATCCfinder\_download

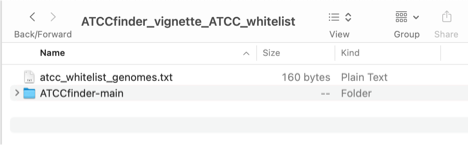
mamba activate ATCCfinder\_download

## Install genome\_portal\_api package

mamba install git pip

pip install /path/to/genome\_portal\_api

1. Ensure all required files and folders are in the same directory (Figure 5), as described in previous steps.



**Figure 5. Example folder directory containing required vignette files and software for running *ATCCfinder* download.py.**

1. Run the following command to download ATCC whitelist reference sequences, specifying your own created API key:

$ ATCCfinder-main/ATCCfinder-main/download.py \

--download reference \

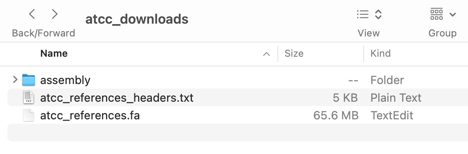
--format True \

--atcc\_ids “atcc\_whitelist\_genomes.txt" \

--out "atcc\_downloads" \

--api\_key <your\_API\_key\_here>

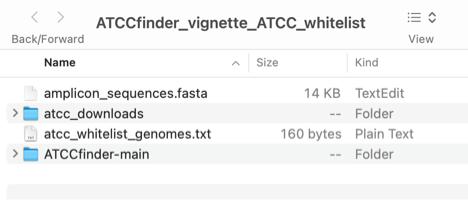
1. Downloading reference genomes may take several minutes, depending on internet speed. Once reference downloading is completed, a new folder “atcc\_downloads” will have been created. Compiled ATCC whitelist references have now been successfully downloaded and compiled into the file “atcc\_downloads/atcc\_references.fa”.



**Figure 6. ATCCfinder folder containing downloaded ATCC sequences and parsed or compiled data.**

Use *ATCCfinder* to query ATCC Whitelist Genomes

1. Ensure all required files and folders are in the same directory (Figure 7), as described in previous steps.



**Figure 7. Example folder directory containing required vignette files and software for running ATCCfinder search.py**.

1. Create and activate a software environment for executing *ATCCfinder* search:

## Define environment

$ mamba create -n ATCCfinder\_search

$ mamba activate ATCCfinder\_search

## Install packages

$ mamba install -c bioconda minimap2

$ mamba install -c bioconda samtools

$ mamba install r-base r-argparse

1. From the command line, navigate to the folder containing vignette files and software. Execute ATCCfinder from the command line:

$ ATCCfinder-main/ATCCfinder-main/search.R \

--query amplicon\_sequences.fasta \

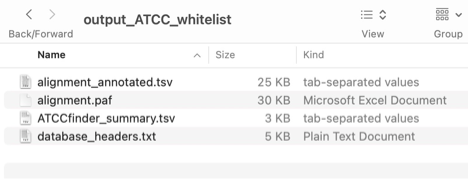
--target atcc\_downloads/atcc\_references.fa \

--nhits 100 \

--overwrite T \

--outdir output\_ATCC\_whitelist

1. Successful execution of the above command will yield a new folder titled “output\_ATCC\_whitelist” containing analysis results (Figure 8). Output file “alignment.paf” contains the raw alignment results generated by searching the query sequence against the custom database, “alignment\_annotated.tsv” contains alignment results with additional meta data annotations and scoring metrics appended to it, and “ATCCfinder\_summary.tsv” provides a result summary for each sequence queried. An example interpretation of the “ATCCfinder\_summary.tsv” file for this analysis vignette is provided in Table 3.



**Figure 8. Results folder after running ATCCfinder search.R.**

The resulting “summary.tsv” file contains analysis result information for each query amplicon sequence, and could be interpreted as follows:

Query sequences report varying qualities of top alignment, as informed by the maximum percentage of aligned bases between query and reference (max\_nmatchh\_p), ranging from 21.9% - 99.8%. We know that we are searching amplicon sequences and whole genome sequences, so only highly similar alignments (max\_nmatchh\_p ≥ 97%) should be considered. Two query sequences, NR\_118997.2 and NR\_114042.1, corresponding to *Staphylococcus aureus* and *Escherichia coli* amplicon sequences, have alignment scores above this threshold and subsequently align against the *Staphylococcus aureus* and *Escherichia coli* ATCC whitelist references. A third query, NR\_119108.1 (*Salmonella enterica*) also reports a relatively strong max\_nmatchh\_p result of 97% against the *Escherichia coli* reference, supporting a close evolutionary relationship between *Salmonella enterica* and *Escherichia coli* and emphasizing the need to be aware of the biological context within which *ATCCfinder* is used so that results may be properly interpreted and false positives avoided.