Contents

[CANDY Documentation and User Guide 2](#_Toc8200867)

[**Overview** 2](#_Toc8200868)

[**Usage (create database)** 2](#_Toc8200869)

[Required arguments 2](#_Toc8200870)

[Optional arguments 2](#_Toc8200871)

[Alternative usage (update annotations in existing database) 2](#_Toc8200872)

[Required arguments 2](#_Toc8200873)

[**Input files** 3](#_Toc8200874)

[**Explanation** 4](#_Toc8200875)

[**Execution** 5](#_Toc8200876)

[Output files (based on the above command) 5](#_Toc8200877)

[**Updating the annotation mapping file** 6](#_Toc8200878)

[**Installation** 6](#_Toc8200879)

# CANDY Documentation and User Guide

## Overview

CANDY provides a command line interface to create customized amplicon databases using *in silico* PCR with the specified primer sequences. CANDY will create database files that serve as input for SMORE'D, a k-mer based classification tool. The required input for this program is a tabular format file containing the primers and their accompanying metadata. Optionally, additional FASTA files containing annotated sequences to include in the generated database can be used. There are two types of sequences that may be added in the additional FASTA files, presence/absence sequences and characterization sequences. These sequences differ in the way they are reported by SMORE'D. SMORE'D reports the number of times a read matches a presence/absence sequence. Only the most abundant of the sample characterization categories is reported *e.g*. , if there are 10,000 *Mycoplasma pneumoniae* P1-type 1 sequences and 1,000 *Mycoplasma pneumoniae* P1-type 2, the sample is characterized as *Mycoplasma pneumoniae* P1-type 1 and only the read count for *Mycoplasma pneumoniae* P1-type 1 is reported. CANDY will output a FASTA file of representative sequences for the database, a profile file which maps the representative sequences to their annotations, and a configuration file.

## Usage (create database)

candy.py –i <primer table> [-o <prefix for output files>] [-d <directory for intermediate files>][-g <additional presence/absence seqs>] [-a <additional characterization seqs>] [-t <num threads>] [-m <mapping file>]

[Bracketed arguments are optional]

### Required arguments

-i, --input Input primer file; tab-separated, 6-column file

### Optional arguments

-o, --output\_prefix Takes a string to be added to the beginning of the output files (amplicons.fasta, profile.tsv, and config.txt)

-g, --pres\_abs\_seqs Takes a FASTA file of presence/absence sequences to be included in the database

-a, --charac\_seqs Takes a FASTA file of sample characterization sequences to be included in the database

-d, --intermediate\_dir Is the name for the directory which holds intermediate files produced. Default name is date-time

-t, --threads Takes an integer argument which tells CANDY how many threads to use

-m, --mapping Takes a tab-separated, 2-column annotation mapping file

-l, --log Takes a name for the log

### Alternative usage (update annotations in existing database)

candy.py --update -p <profile file> -m <mapping file>

### Required arguments

--update Tells CANDY to run in update mode

-p, --profile The existing profile file to be updated

-m, --mapping The updated annotation mapping file

## Input files

*Required:*

***File Description: Input primer file***

* Contains the primers/assays used in targeted sequencing as well as defined metadata
* 6-column, tab-separated file
  + Column 1: Primer name
  + Column 2: Taxonomy ID of target organism
  + Column 3: Target organism
  + Column 4: Virus/bacteria category
  + Column 5: Forward primer sequence
  + Column 6: Reverse primer sequence
* Make sure that no additional spaces are present in cells if creating or editing the file in Excel.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Adenovirus\_1 | 10509 | Adenovirus | Virus | TACATGCACATCGCCG | CCACTGTGGGGTT |
| Adenovirus\_2 | 10509 | Adenovirus | Virus | TACATGCACATCTCGG | CCACCGTGGGGTT |
| A\_baumannii | 470 | A baumannii | Bacteria | CTGCTAATCCAAATCACAG | GGTACATACTCGGTCG |
| B\_brochiseptica | 517 | Bordetella | Bacteria | ATGCACATTTACGGAAATATGA | GCTTGAAGTCATCCAG |

*Optional:*

***File Description: Presence/absence sequences***

* FASTA file with sequences that will be used for presence/absence classifications
* These are sequences identified by methods other than *in silico* PCR such as positive controls, drug resistant/sensitive sequences, or sequences whose variation is not encompassed by NCBI taxonomy (e.g., *Legionella pneumophila* sg1)
* The raw count of reads which match these sequences are reported as the output of SMORE'D
* Each sequence should be labeled with the desired annotation

**>Legionella\_pneumoniae\_sg1**

CTCTGGCTTTGCAGTTATTTTATTACTCCACTCCAGCGATTTACCCTGTTTCTGCTGTGCCTGTGTGGGCTAAACCAGTTAATACCAGTTATG...

**>Mycoplasma\_pneumoniae\_macrolide\_resistant**

TCCGTCCCGCTTGAATGGTGTAACCATCTCTTGACTGTCTCGGCTATAGACTCGGTGAAATCCAGGTACGGGTGAAGATAGGCTGAAGATAGG...

**>H1-275-oseltamivir\_Resistant\_(H275Y)**

CCTCATACAAAATCTTCAGAATAGAAAAGGGAAAGATAATCAAATCAGTCGAAATGAAAGCCCCTAATGAAGATAGGTTATTATTATGATTAT...

**>H3-292-oseltamivir\_Sensitive**

TGCTCCTGCTATCCTCGATATCCTGGTGTCAGATGTGTGAAGATAGGTCTGCAGAGACAACTGGAAAGGATCCAACCGGCCCATCGTAGCGGC...

**>L.\_pneumophila\_sg1\_positive\_control**

CACCTATCAATGCAAAATATCAGGCATGTTCGCTGGAGGTCTGAGGCTGAAGATAGGTGCTGTCGGGGCAAATCTTCCCAGGCGAGAATCTTC...

\*Resistant sequences must come after sensitive sequences when including both sensitive and resistance classifications.

***File Description: Sample characterization sequences***

* FASTA file with sequences that will be used for sample characterization, where only one type is reported per sample
* These are sequences identified by methods other than *in silico* PCR such as *Mycoplasma pneumoniae* P1-Type 1 and *Mycoplasma pneumoniae* P1-Type 2
* Only one type (the most abundant) is reported per sample by SMORE'D
* The sequences must be labeled as **>target shorthand\_variantNumber\_annotation**

**>type\_1\_Mycoplasma pneumoniae P1-type 1**

GGTATAATTGTTTGGATTCGCCATAAATTTTGCTTATGGCTTTTAACTTTCCAACGTGCACATTAAATGGAAGACAATTAAATGGAAGACAATAACAAAACGCAAGCTTACGATTCCAGTAGCATTAAGATTC

**>type\_2\_Mycoplasma pneumoniae P1-type 2**

GAATCTTAATGCTACTGGAATCGTAAGCTTGCGTTTTGTTATTGTCTTCCATTTAATGTGCACGTTGGAAAGTTAAAAGCCATAAGCAAAATTTATGGCGAATCCAAACAATTATACC

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

***File Description: Annotation mapping file***

* Tab-separated, 2-column file that maps the NCBI taxonomy to the desired annotation for reporting
* The first column contains the NCBI taxonomy of the virus at the species level and below
* The second column contains the desired annotation for that taxonomy
* CANDY will replace any instance of the NCBI taxonomy with the desired annotation for sequences in the database

|  |  |
| --- | --- |
| Alpaca adenovirus CA/2007 | Non-human adenovirus |
| Antwerp rhinovirus 98/99 | Rhinovirus |
| Enterovirus A;Coxsackievirus A10 | Enterovirus A |
| Betacoronavirus 1;Human coronavirus OC43 | Human coronavirus OC43 |
| Enterovirus B;Echovirus E31 | Enterovirus B |
| Human mastadenovirus B;Human adenovirus 55 | Human mastadenovirus B |

## Explanation

CANDY has been developed to rapidly generate curated databases containing sequences specific to primer panels and the level of annotation appropriate for the targeted region. CANDY takes assay specific primer sequences, along with information on the intended target of the assay (*e.g.* scientific name, taxonomy ID(s), virus/bacteria category) and performs *in silico* PCR to obtain possible *in vivo* variants of the amplicons. *In silico* PCR is performed using the at minimum BLAST+ 2.8.1 which has "taxonomy aware" databases, greatly improving the accuracy of amplicon generation. The *in silico* PCR uses the forward and reverse primer sequences to BLAST against organism specific sequences and then pulls the region of target sequence spanning the primers. The *in silico* amplicons are then annotated using the NCBI taxonomy. For sequences with appropriate annotations outside the scope of NCBI taxonomy, CANDY will take FASTA files of annotated amplicons and incorporate them into the complete database output. CANDY has been optimized to reduce the amount of manual/human curation required to develop such customized databases. The user is responsible for providing:

**Input file (required)**

1. Primer name
2. Taxonomy ID of the intended target of each primer pair
3. Name of the target organism (not used by the script)
4. The category of bacteria or virus in which the primer target falls
5. Forward and reverse primer sequences

**Additional files (optional)**

1. FASTA file of sequences for presence/absence classification
2. FASTA file of sequences for sample characterization

The files necessary for SMORE'D database building and classification are produced as the output of CANDY, *e.g.* **amplicons.fasta**, **profile.tsv**, and **config.txt**. CANDY provides the option (-o, --output\_prefix) to provide a prefix for these files for each database build, simplifying database versioning. Examples of CANDY output are below.

## Execution

CANDY.py -i primer\_file.txt -o db-Apr17 -g pres\_abs\_seqs.fasta -a character\_seqs.fasta -t 10 -m mapping.txt

Where, CANDY.py calls the program,

-i primer\_file.txt is the input file containing the primers and primer metadata,

-o db-Apr17 is the prefix for output files,

-g pres\_abs\_seqs.fasta is a FASTA file containing amplicons for presence/absence classification,

-a character\_seqs.fasta is a FASTA file containing amplicons for sample characterization,

-t 10 is the number of threads to use, and

-m mapping.txt is the NCBI taxonomy to desired annotation mapping file

### Output files (based on the above command)

***File Description: Config file* (db-Apr17\_config.txt)**

* Text file used by SMORE'D to point to the location of the amplicon and profile files
* Has two sections: loci and profile
* The loci section contains the location of file with the sequences to be detected
* The profile section similarly contains the location of the profile file
* CANDY makes the location of the amplicon and profile file the current working directory (*i.e.* the directory the files will be output to)
* If the amplicon and profile files are moved to a new location, these paths must be updated. (Please use the full path)

[loci]

amplicons /path/to/db-Apr17\_amplicons.fasta

[profile]

profile /path/to/db-Apr17\_profile.tsv

***File Description: Amplicon file* (db-Apr17\_amplicons.fasta)**

* FASTA file containing all of the representative target sequences, this includes amplicons generated via *in silico* PCR as well as the additional sequences provided by the **-g** and/or **-a** options
* Presence/absence sequences are labeled as **>amplicon\_N**. The raw read count will be reported for any sequences that match these amplicons
* Sample characterization sequences will be labeled with **>short hand\_N**. The sequence which matches the most reads per sample will be reported

>amplicon\_1

TACATGCACATCGCCGGACAGGATGCTTCGGAGTACCTGAGTCCGGGTCTGGTGCAGTTCGCCCGTGCAACAGACACCTACTTCAGTATGGGGAACAAGTTTAGAAACCCCACAGTGG

>amplicon\_2

TACATGCACATCGCCGGACAGGATGCTTCGGAGTACCTGAGTCCGGGTCTGGTGCAGTTCGCCCGTGCAACAGACACCTACTTCAGTATGGGGAACAAATTTAGAAACCCCACAGTGG

...

>amplicon\_96

CAGCCAGATGTTTGAATAGCGCGCTGTCGATTTTTTCCAGTTTCACCCACGACTCCACCATGCCTTCGGCACTGCGCACGCCGCTGCGGTTTTCACCGACAACGA

>type\_1

GGTATAATTGTTTGGATTCGCCATAAATTTTGCTTATGGCTTTTAACTTTCCAACGTGCACATTAAATGGAAGACAATTAAATGGAAGACAATAACAAAACGCAA

>type\_2

GAATCTTAATGCTACTGGAATCGTAAGCTTGCGTTTTGTTATTGTCTTCCATTTAATGTGCACGTTGGAAAGTTAAAAGCCATAAGCAAAATTT

***File Description: Profile file* (db-Apr17\_profile.tsv)**

* Tab-separated, 3-column file that matches the sequences in the amplicon file with the desired annotation for that sequence (columns 1 & 2 map to the sequence header in the amplicon file, column 3 is the desired annotation for that sequence)
* Reads with the same annotation (*e.g.* Adenovirus) will be summed together and the total number of reads for that annotation will be reported

amplicon 1 Adenovirus

amplicon 2 Adenovirus

...

amplicon 96 RNaseP\_(human)\_positive\_control

type 1 Mycoplasma pneumoniae P1-type 1

type 2 Mycoplasma pneumoniae P1-type 2

## Updating the annotation mapping file

Because of the diverse nature of taxonomies, the desired annotation for an amplicon (strain/species type or common name) may not agree with the NCBI taxonomy for the target organism. For this reason, CANDY can take as input an annotation mapping file (see ***File Description: Annotation mapping file*** above). This file allows users to provide alternative annotations for any sequences. The default mapping file has been generated, but additions to the file can be made if new primers are added to the database or a new file can be generated. To add taxonomy-to-annotation mappings to the annotation file, users can append the unwanted annotation reported by SMORE'D followed by the replacement annotation (separated by a tab).

You can update an existing profile file by running:

CANDY.py --update -p <profile file> -m <mapping file>

Where,

CANDY.py --update tells CANDY to run in update mode,

-p <profile file> is a previously generated profile file, and

-m <mapping file> is the mapping file with the updated annotations

## Installation

In order to run, CANDY requires the minimum version of **BLAST+ 2.8.1**, its accompanying database (**v5**), **TaxonKit**, **VSEARCH**, and **Python v3** or greaterand thecandy.py script available at [https://github.com/appliedbinf/URDO-CANDy](https://www.google.com/url?q=https://github.com/appliedbinf/URDO-CANDy&sa=D&source=hangouts&ust=1582727853266000&usg=AFQjCNHZpapZ2UGsT1j9lLaR4-bUbiDttA).

**To download BLAST and the corresponding BLAST nucleotide database:**

Download the latest version of [BLAST](ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/) for your system:

wget <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.9.0+-x64-linux.tar.gz>

tar xvf ncbi-blast-2.9.0+-x64-linux.tar.gz

export PATH=$HOME/ncbi-blast-2.9.0+/bin:$PATH

\* export PATH=$HOME/ncbi-blast-2.9.0+/bin:$PATH may be placed in ~/.bashrc

**Download the latest BLAST nucleotide database:**

mkdir $HOME/blastdb #create a directory to store the BLAST database

export BLASTDB=$HOME/blastdb # Set BLASTDB to the directory the database is downloaded to (in .bashrc)

cd $HOME/blastdb

update\_blastdb.pl --decompress v5

\* export BLASTDB=$HOME/blastdb may be placed in ~/.bashrc

**Note:**

Downloading the newest version of BLAST and the v5 database will take time and memory (~64 GB). If these have previously been downloaded you can skip the above steps and place the following commands in your .bashrc file.

export PATH=/path/to/newest/version/of/BLAST/:$PATH

export BLASTDB=/path/to/directory/containing/BLAST/database/v5

**To download** [**TaxonKit**](https://bioinf.shenwei.me/taxonkit/download/) **v0.3.0:**

Download and decompress these files to a directory in your PATH:

<https://github.com/shenwei356/taxonkit/releases/download/v0.3.0/taxonkit_linux_amd64.tar.gz>

tar xvf taxonkit\_linux\_amd64

And copy nodes.dmp and names.dmp to $HOME/.taxonkit

or install via conda:

conda install -c bioconda taxonkit

**Download and install** [**VSEARCH**](https://github.com/torognes/vsearch) **v2.13.3:**

Download the appropriate executable for your system from <https://github.com/torognes/vsearch/releases>

wget <https://github.com/torognes/vsearch/releases/download/v2.13.3/vsearch-2.13.3-linux-x86_64.tar.gz>

tar xvf vsearch-2.13.3-linux-x86\_64.tar.gz

Place /vsearch-2.13.3-linux-x86\_64/bin/ in your PATH.