Contents

[SMORE'D Documentation and User Guide 2](#_Toc33522144)

[Overview 2](#_Toc33522145)

[Usage (Database Building) 2](#_Toc33522146)

[Input Files (Database Building) 3](#_Toc33522147)

* [Output Files for Database Building/ Input Files for Read Classification 4](#_Toc33522148)
* [Explanation 4](#_Toc33522149)
* [Additional Input Files 4](#_Toc33522150)

[Usage (Read Classification) 4](#_Toc33522151)

* [Explanation 5](#_Toc33522152)
* [Output 5](#_Toc33522153)

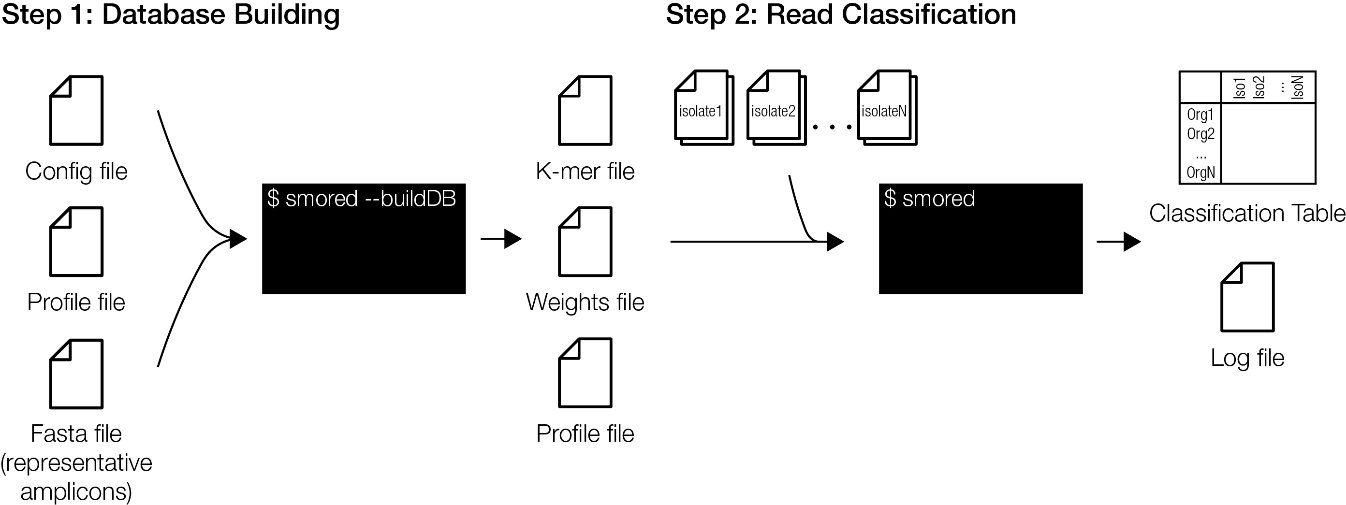
[Example single sample report 7](#_Toc33522156)

[Installation 9](#_Toc33522157)

# SMORE'D Documentation and User Guide

# Overview

**S**equence **M**atching f**O**r **RE**piratory **D**iseases, SMORE'D, is a command-line sequence classification tool tailored to meet the needs of the Undiagnosed Respiratory Disease Outbreak (URDO) branch at CDC. SMORE'D is a k-mer based classification tool capable of rapidly classifying read sequences generated by multi-pathogen detection platforms. These platforms use targeted amplification and whole genome sequencing of bacterial and viral organisms in clinical samples generating datasets of unidentified amplicon sequences. SMORE'D classifies these amplicon sequences at the level of annotation desired for each target-specific assay, whether it be identification or phenotypic characterization. Using a complete and well-curated database of representative target sequences as input, SMORE'D works in two steps. First, SMORE'D builds a k-mer database for the supplied representative target sequences (this is done only once for a given set of target sequences). It then classifies amplicons from paired-end reads and generates a report of all identified targets and the number of reads matching each target. SMORE'D creates an optional single sample report in Excel format that summarizes the sample and provides read counts and relative abundance of all identified organisms. This workflow can be seen in **Figure 1.**



**Figure 1.** SMORE'D workflow. SMORE'D first builds a k-mer database with the necessary input files. SMORE'D then takes paired-end FASTQ files and generates a classification table of identified targets and their read counts and an optional, formatted single sample report.

# Usage (Database Building)

smored --buildDB -c <config file> [-k <int>] [-P|--prefix <database prefix>] [-a <log file path>]

[Bracketed arguments are optional]

Required arguments:

|  |  |
| --- | --- |
| --buildDB | Tells SMORE'D to run in build mode |
| -c, --config | File which points to the locations of necessary inputs (profile file and amplicon file) |

Optional arguments:

|  |  |
| --- | --- |
| -k | Integer number for the k-mer size to use with database |
| -P, --prefix | String to be used as the prefix for the generated database files (K-mer file, weights file, and profile file) |
| -a | File path for the log file |
| -h, --help | Prints the help manual for SMORE'D |

# Input Files (Database Building)

***File Description: Config file***

* 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
* Please use the full path

[loci]

amplicons /path/to/amplicons.fasta

[profile]

profile /path/to/profile.tsv

***File Description: Profile file***

* 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 P-type 1

type 2 Mycoplasma pneumoniae P-type 2

***File Description: Amplicon file***

* FASTA file containing all of the representative target sequences
* 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

## Output Files for Database Building/ Input Files for Read Classification

***File Description: k-mer file***

* Tab-separated, 3-column file which stores all k-mers of length *k* seen in the database and the representative sequence(s) the k-mer can be found in
* Column 1 is the k-mer, column 2 the sequence type, and column 3 is the sequence(s) the k-mer is present in

|  |  |  |
| --- | --- | --- |
| ATTTTTAAAGCTGCTCTTTCTGTAAATGCTTCCCC | amplicon | [1,2] |
| GGGGAAGCATTTACAGAAAGAGCAGCTTTAAAAAT | amplicon | [1,2] |
| TTTTTAAAGCTGCTCTTTCTGTAAATGCTTCCCCA | amplicon | [1,2] |
| AAAGCTGCTCTTTCTGTAAATGCTTCCCCACAATC | amplicon | [1,2,7] |
| GATTGTGGGGAAGCATTTACAGAAAGAGCAGCTTT | amplicon | [1,2,7] |
| TTCTCTCCATTTAGATAATTCATATTTAAAACGTC | amplicon | [1,2,7,8,9] |
| ACGTTTTAAATATGAATTATCTAAATGGAGAGAAG | amplicon | [1,2,7,8,9] |

***File Description: Weights file***

* Tab-separated, 2-column file with each amplicon and its calculated weight
* Not currently being used by SMORE'D

|  |  |
| --- | --- |
| amplicon\_1 | 0.87304 |
| amplicon\_2 | 0.87304 |
| amplicon\_3 | 0.87304 |
| amplicon\_4 | 0.87304 |
| amplicon\_5 | 0.87304 |

***File Description: Profile file***

* The same profile file as the input profile file
* Will be renamed with the output prefix if one is given

## Explanation

SMORE'D uses the amplicon file of representative target sequences to build a k-mer database by breaking the sequences into smaller sequences of size *k* (35bp is the default k-mer size). The k-mer database then stores a record of all k-mers and each sequence a k-mer is found in (refer to *File description: k-mer file*). A weights file stores a calculated weight for each amplicon, calculated by dividing the length of the amplicon by the average length of all amplicons in the database. (This file is currently not used during classification but historically was created to normalize amplicons for read length.) Finally, SMORE'D will make a copy of the input profile file, renaming it with the output prefix, if one is given (-P <prefix>). The generated files (profile file, k-mer file, and weights file) will be used as the database for SMORE'D read classification. SMORE'D only needs to build a database once per set of representative sequences.

## Additional Input Files

In addition to the database files created in database building, SMORE'D now takes two more files that facilitate the creation of the single sample report (example below): amplicon2taxonomy.tsv and template.xlsx. **These are installed when SMORE'D is installed in $PYTHONPATH/lib/smored, and no user input is needed for these.** SMORE'D looks for these files only if the --report flag is provided during read classification. **amplicon2taxonomy.tsv** is a tab-separated file that maps the amplicon names seen in the profile file to a simplified taxonomy. The simplified taxonomy file provides the levels of taxonomy for each amplicon annotation necessary to format the report. **Template.xlsx** is an Excel template used to generate the single sample report.

# Usage (Read Classification)

smored --predict -c <config file> -1 <fwd read FASTQ> -2 <rev read FASTQ> | -d <input directory [-o <output file>] [-P|--prefix <database prefix>] [-k <int>] [-r] [-R <classified output directory>] [-u] [-U <unclassified output directory>][--report] [-x] [-t <num of threads>] [-v] [-h]

Required arguments:

|  |  |  |
| --- | --- | --- |
| -c, --config | | File which points to the locations of necessary inputs (same as for database building) |
| Single  Sample  Mode | -1,  --fastq1 | Path to first FASTQ file for paired-end sample |
| -2,  --fastq2 | Path to second FASTQ file for paired-end sample |
| Multi-  Sample  Mode | -d, --dir | A directory containing paired-end read files for multi-sample classification |

Optional arguments:

|  |  |
| --- | --- |
| --predict | smore'd runs this mode by default |
| -o, --output | File path to save classification output, if nothing is provided the results will print to the screen |
| -P, --prefix | Prefix for the database created. This can include a path to the database file if they are not located in the current working directory |
| -k | K-mer size used to build the database (Default=35) |
| -r | Create FASTA file(s) of the classified reads labeled with their classification and coverage. Deposited in the working directory |
| -R, --readsdir | Path to classified reads output directory for FASTA files used instead of -r |
| -u | Create FASTA file(s) of the unclassified reads. Deposited in the current directory. |
| -U | Path to unclassified reads output directory. |
| --report | Generate per-sample report in Excel format. |
| -x | Overwrites to previous results written to the output file. By default, SMORE'D will append new results to previous results if given the same output file name |
| -t | An integer number of threads to use to process samples |
| -v | Prints the version of the software |
| -h, --help | Print the help manual for this application |

## Explanation

Once the database has been built, SMORE'D is ready to classify sequences from paired-end FASTQ files in either single sample or multi-sample mode. SMORE'D classifications are done using k-mer matching. After merging reads from the FASTQ files (read 1 and read 2), SMORE'D k-merizes the merged reads. Read k-mers are searched against k-mers in the SMORE'D database. The read will be classified as the amplicon in the database with the most matching k-mers. Once all reads and samples have been classified, SMORE'D will generate a read counts table. Each column is a sample, and each row is an identified organism.

## Output

Output files are only produced if the flags or file paths are provided in the command. Otherwise, a table with the identified organisms and their counts are printed on the screen.

## 

***File Description: Output file***

* Created if -o <file name> or --output <file name> is used
* Tab-separated file
* The first column lists all identified organisms for the sample(s) and the following columns report the number of reads classified as that organism in each sample
* Rows correspond to different identified organisms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample | MID-104 | MID-10 | MID-11 | MID-124 | MID-12 |
| Chlamydia abortus | 14551 | 1 | 3 | 1 | 1 |
| Chlamydia psittaci | 29 | 6 | 12 | 0 | 13 |
| Streptococcus pyogenes | 2 | 61 | 262216 | 0 | 2332 |
| Bordetella pertussis | 1 | 13 | 25 | 0 | 30 |
| Streptococcus pneumoniae | 3 | 1061 | 329 | 0 | 189 |
| Chlamydia pneumoniae | 8 | 21 | 35 | 0 | 38 |
| Streptococcus mitis | 1 | 93 | 130 | 1 | 101 |
| Bordetella parapertussis | 1 | 4 | 4 | 0 | 2 |

## 

***File Description: Classified reads***

* If -r is used, FASTA file(s) of classified reads are placed in the working directory
* If -R <directory name> is used, classified reads are placed in the provided directory
* FASTA files containing classified read sequences
* The semicolon-separated header for each read sequence contains the original read identifier, the number of identical reads before deduplication, the read k-mer coverage, and the organism the read is classified as

>M04906:103:000000000-BVL7N:1:1101:13607:1813;size=101218;0.855;Streptococcus pneumoniae

CCATCTGCATTCATAAAGTGGTAGTAGTTTGGTTCCAAATCTTCAAAGAGATCGACTTTGGCTGTGTGGTTATAAACGACATCTAGGATAGCTCCCATACCACGTTTGTG

>M04906:103:000000000-BVL7N:1:1101:18925:1815;size=82028;0.855;Streptococcus pneumoniae

CCATCGGCATTCATAAAGTGGTAGTAGTTTGGTTCCAAATCTTCAAAGAGATCGACTTTGGCTGTGTGGTTATAAACGACATCTAGGATAGCTCCCATACCACGTTTGTG

***File Description: Unclassified reads***

* Created if -u is used, places the FASTA file(s) of unclassified reads in the working directory
* -U <directory name> is used unclassified reads are placed in the provided directory
* FASTA files of unclassified read sequences
* The semicolon-separated header for each read sequence contains the original read identifier, the number of identical reads before deduplication, and unclassified

>M04906:103:000000000-BVL7N:1:1101:22837:11881;size=103;unclassified

CCATCTGCATTCATAAAGTGGTAGTAGTTTGGTTCCAAACCTTCAAAGAGATCGACTTTGGCTGTGTGGTTATAAACGGCATCTAGGATAGCTCCCATACCACGTTTGTG

>M04906:103:000000000-BVL7N:1:1101:10178:10602;size=92;unclassified

CCATCTGCATTCATAAAGTGGTAGTAGTTTGGTTCCAAACCTTCAAAGAGATCGACTTTGGCTGTGTGGTTATAAACGGCATCTAGGATAGCTCCCATGCCACGTTTGTG

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Example single sample report | |  |  |  | |  | | --- | |  | |
| **CDC SMORE'D Classification Report** | | | | | |
| **Sample:** | 10 |  |  |  |  |
| **Date:** | 2019-12-18 |  |  |  |  |
| **Total # of reads sequences:** | 412365 |  |  |  |  |
| **Total # of merged reads:** | 206607 |  |  |  |  |
| **Most abundant organisms** | | | | | |
| **Name** | **Genus** | **Species** | **Type / Resistance** | **% of reads** |  |
| *Haemophilus influenzae* | *Haemophilus* | *Haemophilus influenzae* |  | 18.18 |  |
| *Human RNaseP* |  |  |  | 15.53 |  |
| *Influenza A virus* | *Alphainfluenzavirus* | *Influenza A virus* |  | 15.26 |  |
|  |  |  |  |  |  |
| **Bacteria** | | | | | |
| **Name** | **Genus** | **Species** | **Type / Resistance** | **# of reads** | **% of reads** |
| *23S-Ureaplasma parvum Sensitive* | *Ureaplasma* | *parvum* | Sensitive | 1 | 0.00 |
| *Acinetobacter baumannii* | *Acinetobacter* | *baumannii* |  | 1 | 0.00 |
| *Bordetella holmesii* | *Bordetella* | *holmesii* |  | 1 | 0.00 |
| *Bordetella pertussis* | *Bordetella* | *pertussis* |  | 2 | 0.00 |
| *Chlamydia pneumoniae* | *Chlamydia* | *pneumoniae* |  | 3 | 0.00 |
| *Haemophilus influenzae* | *Haemophilus* | *influenzae* |  | 37553 | 18.18 |
| *Klebsiella pneumoniae* | *Klebsiella* | *pneumoniae* |  | 4 | 0.00 |
| *Legionella pneumophila sg1* | *Legionella* | *pneumophila* |  | 4 | 0.00 |
| *Moraxella catarrhalis* | *Moraxella* | *catarrhalis* |  | 16364 | 7.92 |
| *Pseudomonas aeruginosa* | *Pseudomonas* | *aeruginosa* |  | 1 | 0.00 |
| *Staphylococcus aureus* | *Staphylococcus* | *aureus* |  | 9 | 0.00 |
| *Streptococcus agalactiae* | *Streptococcus* | *agalactiae* |  | 7 | 0.00 |
| *Streptococcus mitis* | *Streptococcus* | *mitis* |  | 11830 | 5.73 |
| *Streptococcus oralis* | *Streptococcus* | *oralis* |  | 7 | 0.00 |
| *Streptococcus pneumoniae* | *Streptococcus* | *pneumoniae* |  | 5109 | 2.47 |
| *Streptococcus pseudopneumoniae* | *Streptococcus* | *pseudopneumoniae* |  | 10233 | 4.95 |
| *Streptococcus sp. ChDC B345* | *Streptococcus* | *sp. ChDC B345* |  | 152 | 0.07 |
| *Streptococcus sp. NPS 308* | *Streptococcus* | *sp. NPS 308* |  | 114 | 0.06 |
|  |  |  |  |  |  |
| **Virus** | | | | | |
| **Name** | **Family** | **Species** | **Drug resistance** | **# of reads** | **% of reads** |
| *Enterovirus* | Picornaviridae |  |  | 6 | 0.00 |
| *Enterovirus C* | Picornaviridae | *Enterovirus C* |  | 1 | 0.00 |
| *Enterovirus D* | Picornaviridae | *Enterovirus D* |  | 22 | 0.01 |
| *H1-275-oseltamivir Sensitive* | Orthomyxoviridae | *Influenza A virus* | H1-275-oseltamivir Sensitive | 3 | 0.00 |
| *H3-119-oseltamivir Resistant (I122V)* | Orthomyxoviridae | *Influenza A virus* | H3-119-oseltamivir Resistant (I122V) | 28787 | 13.93 |
| *H3-119-oseltamivir Sensitive* | Orthomyxoviridae | *Influenza A virus* | H3-119-oseltamivir Sensitive | 1132 | 0.55 |
| *H3-292-oseltamivir Sensitive* | Orthomyxoviridae | *Influenza A virus* | H3-292-oseltamivir Sensitive | 1 | 0.00 |
| *HPIV3* | Paramyxoviridae | *Human respirovirus 3* |  | 1 | 0.00 |
| *Human respiratory syncytial virus A* | Pneumoviridae | *Human orthopneumovirus* | | 1 | 0.00 |
| *Influenza A virus* | Orthomyxoviridae | *Influenza A virus* |  | 31527 | 15.26 |
| *Rhinovirus C* | Picornaviridae | *Rhinovirus C* |  | 11 | 0.01 |
|  |  |  |  |  |  |
| **Other** |  |  |  |  |  |
| **Name** |  |  |  | **# of reads** | **% of reads** |
| Human RNaseP |  |  |  | 32083 | 15.53 |
| Positive control HBoV |  |  |  | 2 | 0.00 |
| Positive control HCoV-1 |  |  |  | 2 | 0.00 |
| Positive control HCoV-2 |  |  |  | 1 | 0.00 |
| Positive control HPEV |  |  |  | 1 | 0.00 |
| Positive control HPIV2 |  |  |  | 3 | 0.00 |
| Positive control HPIV4 |  |  |  | 1 | 0.00 |
| Positive control M. catarrhalis |  |  |  | 4 | 0.00 |
| Positive control M. pneumoniae - P1 typing |  |  |  | 5 | 0.00 |
| Positive control M. pneumoniae - macrolide resistance | |  |  | 6 | 0.00 |
| Positive control Pan-FluA |  |  |  | 1 | 0.00 |
| Positive control Pan-Legionella |  |  |  | 2 | 0.00 |
| Positive control RNaseP (human) |  |  |  | 1 | 0.00 |

# Installation

SMORE'D can be found at [https://github.com/appliedbinf/URDO-SMOREd/releases](https://www.google.com/url?q=https://github.com/appliedbinf/URDO-SMOREd/releases&sa=D&source=hangouts&ust=1582732694077000&usg=AFQjCNF2XwqxS_w6e7iTmkHzUirSBZDFWw).

To install:

wget https://github.com/appliedbinf/URDO-SMOREd/releases/download/1.0/SMOREd-1.0-Linux-x86\_64.sh

chmod +x SMOREd-1.0-Linux-x86\_64.sh

./ SMOREd-1.0-Linux-x86\_64.sh

Smore'd will provide prompts to direct the remaining installation.