

NAME

vsearch — dereplicate, filter, sort, search, compare and clusterize amplicons from metagenomic projects

SYNOPSIS

vsearch [*options*] *filename*

DESCRIPTION

Environmental or clinical molecular studies generate large volumes of amplicons (e.g. SSU-rRNA sequences) that need to be filtered, dereplicated, searched, clustered or compared to sequences from other studies. The aim of **vsearch** is to offer a all-in-one open source tool to perform these tasks, using optimized algorithm implementations and harvesting the full potential of modern computers to guarantee the fastest and more accurate possible processing.

Nucleotidic sequence comparisons is at the core of **vsearch**. To speed up comparisons, **vsearch** implements *k*-mer filtering, and an extremely fast Needleman-Wunsch algorithm making use of the Streaming SIMD Extensions (SSE2) of modern x86-64 CPUs. If SSE2 instructions are not available, **vsearch** exits with an error message.

vsearch input is a fasta file containing one or several nucleotidic sequences. For each sequence, the sequence identifier is defined as the string comprised between the ">" symbol and the first space or the end of the line, whichever comes first. Additionally, if the line ends with the pattern ";size=*integer*;", **vsearch** will interpret *integer* as the abundance of the sequence (in a dereplicated fasta file for instance). The nucleotidic sequence is defined as a string of [acgt] or [acgu] symbols (case insensitive), starting after the end of the identifier line and ending before the next identifier line or the file end; **vsearch** exits with an error message if any other symbol is present in the sequence. Optionally, **vsearch** can be compiled to accepted compressed fasta files as input (gz and bzip2 formats).

Options

vsearch recognizes a large number of command-line options. For an easier navigation, options are grouped by theme (dereplication, filtering, sorting, searching, comparison, clustering). We start with general options that apply to all themes.

General options:

- help** display a short help and exit.
- version** output version information and exit.
- fasta_width** *positive integer*
fasta files produced by **vsearch** are wrapped (sequences written on lines of *integer* nucleotides, 80 by default). Set that value to 0 to eliminate the wrapping.
- maxseqlength** *positive integer*
all **vsearch** operations will discard sequences of length equal or greater than *integer* (50,000 nucleotides by default).
- minseqlength** *positive integer*
all **vsearch** operations will discard sequences of length smaller than *integer* (1 nucleotide by default for sorting or shuffling, 32 nucleotides for dereplication or searching).
- notrunclabels**
do not truncate sequence labels at first space, use the full header.
- strand** *plus/both*
when searching or dereplicating, check the *plus* strand only (default) or check *both* strands.
- threads** *positive integer*
number of computation threads to use. The number of threads should be lesser or equal to the number of available CPU cores. Default number of threads is 1, use 0 to launch a number of threads equal to the number of CPU cores.

--uc *filename*

when searching or dereplicating, output results in *filename* using a uclust-like format.

--uc_allhits

when searching or dereplicating, and when using the uclust-like format option, show all hits, not just top hit.

Clustering options: (in progress)

EXAMPLES

(in progress)

LIMITATIONS

vsearch does not perform chimera detection..

AUTHORS

Implementation by Torbjørn Rognes, documentation by Frédéric Mahé, .

REPORTING BUGS

Submit suggestions and bug-reports at <<https://github.com/torognes/vsearch/issues>>, send a pull request on <<https://github.com/torognes/vsearch>>, or compose a friendly or curmudgeont e-mail to Torbjørn Rognes <torognes@ifi.uio.no>.

AVAILABILITY

The software is available from <<https://github.com/torognes/vsearch>>

COPYRIGHT

Copyright (C) 2014 Torbjørn Rognes

This program is free software: you can redistribute it and/or modify it under the terms of the GNU Affero General Public License as published by the Free Software Foundation, either version 3 of the License, or any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Affero General Public License for more details.

You should have received a copy of the GNU Affero General Public License along with this program. If not, see <<http://www.gnu.org/licenses/>>.

vsearch includes code from Google's CityHash project by Geoff Pike and Jyrki Alakuijala, providing some excellent hash functions available under a MIT license.

vsearch includes code derived from Tatusov and Lipman's DUST program that is in the public domain.

vsearch binaries may include code from the zlib library copyright Jean-loup Gailly and Mark Adler.

vsearch binaries may include code from the bzip2 library copyright Julian R. Seward.

SEE ALSO

swipe, an extremely fast Smith-Waterman database search tool by Torbjørn Rognes (available from <<https://github.com/torognes/swipe>>).

VERSION HISTORY

New features and important modifications of **vsearch** (short lived or minor bug releases are not mentioned):

v1.0 released November 1st, 2014

First public release