

# User's guide: Manual for V-RevComp 1.1

This is a guide to install and use the software utility V-RevComp. The software is reasonably platform-independent. The instructions below should work fine with little or no modifications for nearly all UNIX-type systems including MacOS X. Windows users are advised to install Cygwin (<http://www.cygwin.com>) in order to compile and run the utilities.

- 1. Detailed installation instructions**
- 2. Usage and commands**
- 3. Interpretation of output**
- 4. Discussion**

## 1. Detailed installation instructions

The REAME.txt file bundled with the script provides a quick installation guide.

- 1.1 In order to install certain packages, you might need to have superuser privileges. For installation on Mac, you will have to install the Apple Xcode package available on your MacOS X System DVD in order to be able to compile programs. Please talk to your system administrator if you feel unsure about these steps. Note that the packages are mandatory and that you should not proceed unless these criteria are fulfilled.
- 1.2 Perl needs to be installed on the computer. Most Unix-based systems, including Linux and MacOS X, have Perl pre-installed. You can check this by opening a command line terminal and type “perl -v”. In case Perl is not installed you have to download (<http://www.perl.org>) and compile the program. Windows users can compile Perl with Cygwin or, alternatively, install ActivePerl for Windows (<http://www.activestate.com/activeperl/>).

- 1.3 Download and install HMMER version 3 (<http://hmmer.janelia.org/#download>)

Download the HMMER package version 3 for your operating system to your preferred directory such as /home/user/. Open a command line terminal, move into the directory with “cd /home/user/” and unpack the tarball with “tar xvfz hmmer-3.0.tar.gz”. Enter the new directory and follow the installation instructions in the file INSTALL. The HMMER package should have been compiled and installed on your computer; you can check this by typing “hmmsearch -h” in the terminal and press enter; you should now see HMMER output.

- 1.4 Download V-RevComp

Go to <http://www.cmde.science.ubc.ca/mohn/software.html> in order to download the program and save it to your directory /home/users/. Extract the zip file with “unzip vrevcomp.zip”. A folder called vrevcomp will be created. You will see the following files and directories, vrevcomp.pl (the actual script, including the GPL license), the HMMs directory (containing the Hidden Markov Models), the User's Guide, the README.txt file as well as test input and output files. Go to the vrevcomp/ directory and type “perl vrevcomp.pl”; the command options should be printed on the screen (Figure 2), showing that the installation was successful.

## 2. Usage and commands:

Go to the directory `/home/users/vrevcomp/` and type “`perl vrevcomp.pl`”. This will print all command options on the screen (Figure 2). All options and specifications are also listed in Table 1. You can use the `testfileSSUbact.fasta` that comes bundled with the script for a test run. This file contains 200 randomly selected entries of the SILVA bacterial SSU reference alignment, with 20 sequences deliberately reverse complemented for the purpose of testing. The two output files generated by the script are also included for your convenience.

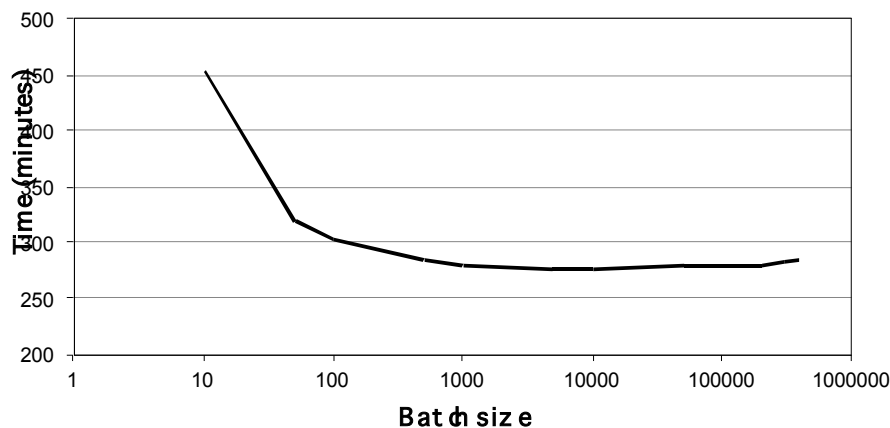
Table 1. List of commands for V-RevComp.

Commands	Description
<b>input file</b>	Name of input FASTA file.
<b>-o</b>	Name of output FASTA file. Sequences that were successfully detected as being in the reverse complementary orientation are reoriented.
<b>-h</b> HMM/SSU/bacteria HMM/SSU/archaea HMM/SSU/fungi HMM/LSU/bacteria HMM/LSU/archaea HMM/LSU/fungi	Name of directory for Hidden Markov Models, e.g. HMMs/SSU/bacteria. The user can specify new directories for other HMMs, e.g. for checking orientation of other targets than SSU or LSU rRNA genes. Files are named <ul style="list-style-type: none"> <li>• <code>V[1-x]leftlong.HMM</code> and <code>V[1-x]rightlong.HMM</code></li> </ul> Note: The program is general enough that as long as the suffix and extension are present, the name of each region is unimportant.
<b>-c</b>	Name of output table containing information about orientation of the query sequences as well as HMM detection frequency on the input and reverse complementary sequence. Orientation is defined as one of the following four categories: <ul style="list-style-type: none"> <li>• forward (query in correct orientation)</li> <li>• reverse (query in reverse complementary orientation)</li> <li>• not found (no HMMs detected in query)</li> <li>• uncertain - forward/reverse/eitherway (HMMs detected in both orientations, detection ratio determines the most likely orientation).</li> </ul>
<b>-r</b> e.g. V1,V2,V3	Specifies whether only specific target regions should be taken into account (e.g. if the user knows which gene segment is screened). By default, all conserved regions flanking the hypervariable regions are considered (except region downstream of V9 for screening SSU sequences, see option <code>-i</code> for more details). Selecting only the regions that are part of the query sequence increases speed considerably.
<b>-i</b>	For SSU rRNA genes, the HMM downstream of V9 is not screened by default, since it features a shorter target sequence and has a slightly weaker performance than the other 17 HMMs. The user can include this particular HMM ( <code>V9rightlong.HMM</code> ) by selecting the <code>-i</code> option.
<b>-b</b>	<i>hmmsearch</i> can use only one threshold parameter, i.e. e-value or score. By default, <i>vxtractor.pl</i> uses the e-value. The user can switch to using the score by selecting this function.
<b>-e</b> <b>-s</b>	Sets the <i>hmmsearch</i> default threshold for e-value ( <code>-e</code> ) and score ( <code>-s</code> ), for HMM files that do not have these values set internally. The bundled HMM files have these values hardcoded (see line starting with “DESC”). The internally set values were tailored for optimal performance. Changing thresholds is useful when testing new HMMs not containing tailored thresholds.
<b>-z</b>	Determines the size of the sequence batch that has to be stepwise processed by V-RevComp. To have larger batches increases speed but also require more memory. Smaller batches are easier on the memory but reduce speed. The default is 1000, but the user can adjust the value based on the memory available. Figure 1 gives a guideline and shows that speed improvement above 1000 is probably insignificant.

**Speed** becomes important when the files contain high numbers of sequences. There are two approaches to increase speed. The two options can be applied in parallel.

**Option1:** Restrict the screen to sub-regions of the rRNA gene sequence. If the user knows which hypervariable sub-regions are represented in the dataset, e.g. if the datasets was generated by amplicon sequencing, the `-r` option (separating the hypervariable regions by a comma such as for example V1,V2,V3 for SSU or V08,V09,V10,V11 for LSU) allows to screen only the HMMs tailored for these particular regions. Cutting down the number of HMMs to be screened significantly improves speed and decreases the chances of false positives.

**Option 2:** Increase the batch size (`-z`). The batch size determines the number of sequences that are screened by HMMER during each run increment. The maximum batch size is the total number of query sequences in the input file. We did not observe significant speed improvements with batch sizes above 1000 (Fig. 1), which was therefore set as the default and should work fine for most datasets.



**Fig. 1.** Speed improvement (i.e. time required to process the dataset) as a function of batch size selected in V-RevComp (option `-z`). A total of 391,178 bacterial sequences deposited in the SILVA database release 102 (<http://www.arb-silva.de>) were processed using different batch sizes. The data were processed on a dual core 3 GHz processor computer.

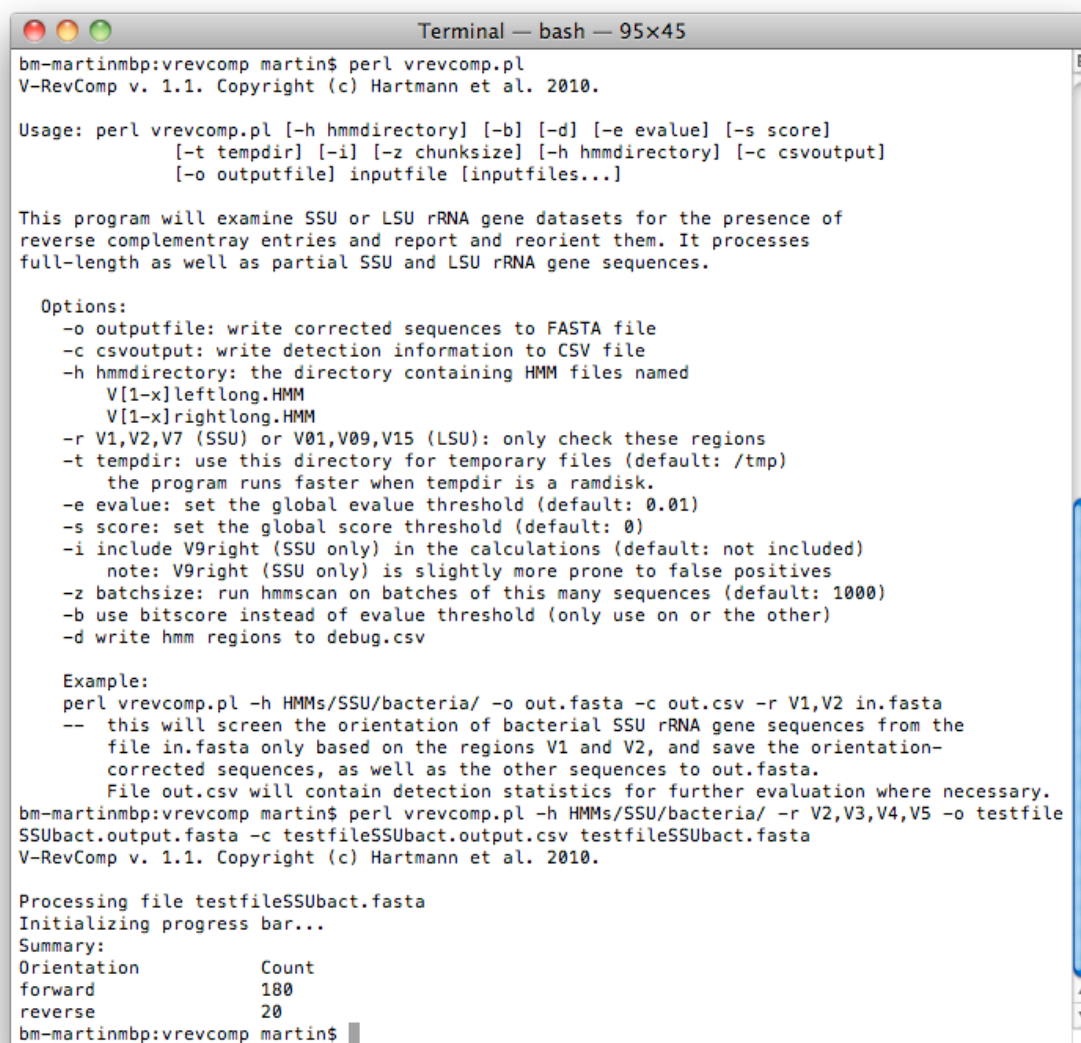
### 3. Interpretation of the output

A progress bar on the screen indicates when the process is expected to be finished (see bottom of screenshot Figure 2). Once the execution of the script has been completed, one or two files will be found in the current directory, i.e. the FASTA and/or comma-separated value (csv) file.

**The FASTA file** contains all input sequences with the sequences determined as reverse complementary being reoriented. No decision can be made for sequences where no HMMs were detected or where HMMs were detected but in equal ratios between original and reverse complementary orientation. These sequences are left in the original orientation.

**The csv file** indicates the orientation of the sequences and reports the number of HMMs detected in the forward (original) and reverse complementary orientation. The decisions reached are termed as “forward” (all HMMs detected in original orientation), “reverse” (all HMMs detected in reverse complementary orientation), “uncertain-forward” (more than 50% of the total HMM detections in the original orientation), “uncertain-reverse” (more than 50% of the total HMM detections in the reverse complementary orientation), “uncertain-eitherway”

(equal HMM counts in both orientations), and “notfound” (no HMMs detected in neither orientation).



```

Terminal — bash — 95x45
bm-martinmbp:vrevcomp martin$ perl vrevcomp.pl
V-RevComp v. 1.1. Copyright (c) Hartmann et al. 2010.

Usage: perl vrevcomp.pl [-h hmmdirectory] [-b] [-d] [-e evaluel] [-s score]
                        [-t tempdir] [-i] [-z chunksize] [-h hmmdirectory] [-c csvoutput]
                        [-o outputfile] inputfile [inputfiles...]

This program will examine SSU or LSU rRNA gene datasets for the presence of
reverse complementray entries and report and reorient them. It processes
full-length as well as partial SSU and LSU rRNA gene sequences.

Options:
-o outputfile: write corrected sequences to FASTA file
-c csvoutput: write detection information to CSV file
-h hmmdirectory: the directory containing HMM files named
    V[1-x]leftlong.HMM
    V[1-x]rightlong.HMM
-r V1,V2,V7 (SSU) or V01,V09,V15 (LSU): only check these regions
-t tempdir: use this directory for temporary files (default: /tmp)
    the program runs faster when tempdir is a ramdisk.
-e evaluel: set the global evaluel threshold (default: 0.01)
-s score: set the global score threshold (default: 0)
-i include V9right (SSU only) in the calculations (default: not included)
    note: V9right (SSU only) is slightly more prone to false positives
-z batchsize: run hmmscan on batches of this many sequences (default: 1000)
-b use bitscore instead of evaluel threshold (only use on or the other)
-d write hmm regions to debug.csv

Example:
perl vrevcomp.pl -h HMMs/SSU/bacteria/ -o out.fasta -c out.csv -r V1,V2 in.fasta
-- this will screen the orientation of bacterial SSU rRNA gene sequences from the
    file in.fasta only based on the regions V1 and V2, and save the orientation-
    corrected sequences, as well as the other sequences to out.fasta.
    File out.csv will contain detection statistics for further evaluation where necessary.
bm-martinmbp:vrevcomp martin$ perl vrevcomp.pl -h HMMs/SSU/bacteria/ -r V2,V3,V4,V5 -o testfile
SSUbact.output.fasta -c testfileSSUbact.output.csv testfileSSUbact.fasta
V-RevComp v. 1.1. Copyright (c) Hartmann et al. 2010.

Processing file testfileSSUbact.fasta
Initializing progress bar...
Summary:
Orientation      Count
forward          180
reverse          20
bm-martinmbp:vrevcomp martin$

```

**Fig. 2.** Screenshot of V-RevComp. Typing “perl vrevcomp.pl” lists all command options on the screen. The command at the bottom shows an actual run of the script using testfileSSUbact.fasta as input. The commands indicate that sequences are processed using bacteria-specific, SSU-specific HMMs (-h HMMs/SSU/bacteria/), only the conserved regions flanking V2, V3, V4, and V5 are considered, and output files testfileSSUbact.output.csv (-c) and testfileSSUbact.output.fasta (-o) will be generated. The run summary at the bottom indicates in this case that 20 out of the 200 sequences were detected as being in the reverse complementary orientation; these sequences are given as re-oriented in the FASTA output file.

## 4. Discussion

Although the HMMs were tailored to be as unforgiving as possible, it is not a good idea to use large sequence information up-or downstream of the SSU and LSU rRNA genes. This increases the chances of false-positive detection. The HMMs will also not perform very well on sequences of poor read quality and many IUPAC ambiguities.

The function *hmmscan* uses heuristic filters to increase the search speed. This reduces the power to detect divergent regions. V-RevComp uses *hmmscan --max* in order to turn off all

heuristic filters. This increases detection power but might reduce speed. If speed is preferable over power, the user might want to adjust the script and remove the `--max` option. However, collecting benchmarks on large datasets we found that speed is not an issue.

V-RevComp cannot be expected to do a good job on partial rRNA gene sequences shorter than 150 bp. At least one flanking region to at least one V region must be present in the sequence data for V-RevComp to process the sequence properly.

If you find that the regions are not located when in fact they should have been, you may need to adjust the *hmmscan* e-values and scores to be more allowing. The tailored e-values and scores were hardcoded in the respective HMM files in the line starting with DESC. However, we do not recommend modifying these values without good reasons, as many trials went into selecting them. In order to find alternate thresholds for your sequences, run *hmmscan --max* by hand for a few of your sequences to see at what e-values and scores the HMMs in question are found, and then change the files accordingly. Note that you may still want the e-values and scores to be as unforgiving as possible in the interest of a low number of false-positive detections. We ask the user to understand that we have made every effort to ensure optimal performance of the HMMs datasets. However, as new entries fill the public databases even better reference alignments can be used to design general HMMs.

If you have large input files – from say pyrosequencing runs – it may be a good idea to have the software process the file in batches, which requires substantially less memory. The switch “-z 1000” will mean that 1000 sequences will be loaded and processed at the time; thus you could process files of arbitrary size with V-RevComp. This pertains only to the way the input data are loaded and does not have any effect on the results of the analysis. We did not see speed improvements with batch sizes bigger than 1000, which is the default setting of V-RevComp.

For the fungal LSU, the HMM pair called V01 is the start of the LSU, V02 corresponds to the formal D1 region, V03 to D2, and so on to V14, which corresponds to D12 (since there are two D7: D7a and D7b; see Nucleic Acids Res. 12 (8), 3563-3583 (1984)). V15 corresponds to the end of the LSU. Thus, V01 and V15 are not formal D regions; they are provided for anyone who also wishes to address the very start, and the very end, of the LSU.