Running VirSorter (Roux et al. 2015) on an HPC Cluster using SLURM

Erica Holdridge

Updated: March 24, 2021

Article: https://peerj.com/articles/985/

GitHub: https://github.com/simroux/VirSorter

1. After opening a terminal window and logging on to your HPC, create a conda virtual environment by running (change "myenv" to the name you would like to call your environment}:

   ```
   conda create --name myenv
   ```

2. Activate the environment by running:

   ```
   conda activate myenv
   ```

3. Download the databases that VirSorter needs:

   ```
   wget https://zenodo.org/record/1168727/files/virsorter-data-v2.tar.gz
   md5sum virsorter-data-v2.tar.gz
   tar -xvzf virsorter-data-v2.tar.gz
   ```

4. Install VirSorter and its dependencies by running:

   ```
   conda install virsorter -c bioconda mcl=14.137 muscle blast perl-bioperl perl-file-which hmmer=3.1b2 perl-parallel-forkmanager perl-list-moreutils diamond=0.9.14
   git clone https://github.com/simroux/VirSorter.git
   cd VirSorter/Scripts
   make clean
   make
   ```

5. Create symbolic links to important files/directories so you can run VirSorter from any directory:

   ```
   ln -s ~/Applications/VirSorter/wrapper_phage_contigs_sorter_iPlant.pl ~/miniconda/envs/virsorter/bin
   ln -s ~/Applications/VirSorter/Scripts ~/miniconda/envs/virsorter/bin
   ```

6. Install MetaGeneAnnotator:

   ```
   conda install --name virsorter -c bioconda metagene_annotator
   ```

7. VirSorter may run into some problems finding files that it needs. To avoid this, run the following commands:

    cpan File::Which
    cpan Parallel::ForkManager

    See if the Bio/ directory is already in a path that @INC can find by running:

    perl -e 'use Bio::Seq'

    If it doesn't return output, you're all set. If it returns something like:

    "Can't locate Bio/Seq.pm in @INC (you may need to install the Bio::Seq module)
    (@INC contains: /bsuhome/eholdridge/perl5/perlbrew/perls/perl-
    5.28.0/lib/site_perl/5.28.0/x86_64-linux
    /bsuhome/eholdridge/perl5/perlbrew/perls/perl-5.28.0/lib/site_perl/5.28.0
    /bsuhome/eholdridge/perl5/perlbrew/perls/perl-5.28.0/lib/5.28.0/x86_64-linux
    /bsuhome/eholdridge/perl5/perlbrew/perls/perl-5.28.0/lib/5.28.0)"

    You need to copy the directory Bio/ into one of the directories listed in the error. For example:

    cp -r miniconda3/envs/virsorter/lib/perl5/site_perl/5.22.0/Bio/
    /bsuhome/eholdridge/perl5/perlbrew/perls/perl-
    5.28.0/lib/site_perl/5.28.0/x86_64-linux

8. The general way to run VirSorter is to activate your conda virtual environment and then run the following command:

    wrapper_phage_contigs_sorter_iPlant.pl -f assembly.fasta --db 1 --wdir
    output_directory --ncpu 4 --data-dir /path/to/virsorter-data

    Where "assembly.fasta" is our input fasta file, db is 1 or 2 depending on which database you want to use (see Roux et al. 2015), "output_dircetory" is where you would like the output to go (note that this should NOT be a folder that already exists), and "/path/to/virsorter-data" is the path that points to the directory where "virsorter-data" was installed.

9. We want to run this on the cluster, likely for a whole bunch of fasta files, so we need a shell script in the following format (for SLURM):

    #!/bin/bash

    #SBATCH --job-name=VirSorter_Microcosm_Metagenomes

```
#SBATCH --ntasks=1
#SBATCH --ntasks-per-node=4
#SBATCH --partition=bsudfq
#SBATCH --mail-type=ALL
#SBATCH --time=24:00:00
#SBATCH --array=1-10
#SBATCH --output=VirSorter_Microcosm_Metagenomes

source /bsuhome/eholdridge/.bashrc
source activate myenv

META_NUM='printf M%02d $SLURM_ARRAY_TASK_ID'

VirSorter/wrapper_phage_contigs_sorter_iPlant.pl -f fasta_folder
/$META_NUM-megahit-contigs.fasta --db 1 --wdir
/bsuhome/eholdridge/virsorter_output/$META_NUM_output/ --ncpu 4 --data-
dir /bsuhome/eholdridge/virsorter-data
```

In this example, the script takes all 10 fasta files in the “fasta_folder” (which should be imported to your HPC account home directory), runs VirSorter on each, and place the output in the directory “virsorter_output” within which each fasta file will have its own output folder. Save this script as a plain text file (“Format” > “Make Plain Text”) and be sure to unlock it (right click > “Get Info” > double click the lock icon at the bottom and enter your password) before copying it to your HPC account home directory.

The fasta files in this case are named “M01-megahit-contigs.fasta” through “M10…”.More info about submitting SLURM job arrays here: https://rc.byu.edu/wiki/index.php?page=How+do+I+submit+a+large+number+of+very+similar+jobs%3F

10. Make the file executable by running (change "shellscript.sh" to the name of your file):

```
chmod u+x shellscript.sh
```

11. Submit the job to SLURM with:

```
sbatch shellscript.sh
```

12. You can keep an eye on the run by checking the error log (change “${base}” to whatever your file names are):

```
Less virsorter_output/${base}_output/logs/err
```