**VirHosFilt**

**Introduction**

VirHosFilt is a software that can be used to remove the host reads from a fastq file produced by the resequencing a viral genome extracted by the host cell. A typical scenario where VirHosFilt could be handy is when a virus is extracted from a clinical sample and the genome from the host can be highly represented within the reads produced by sequencing such a sample.

VirHosFilt allows the user to select among a list of possible hosts and as well as a list of possible virus. The fastq file that needs to be purged by the host reads is aligned against the host genome and, if desired, against a reference genome of the virus.

New host and virus genomes can be added locally or be asked to be added by the software administrator.

After removing the host reads, the purged dataset can be used as an input file for a Kraken analysis in order to highlight all the possible taxon that are present in the sample that can be due to co-infection, contamination etc.

VirHosFilt will output the fastq files of the purged reads, the alignment file on the host and, If requested, on the virus, and the result of the kraken analysis.

**Installation**

If you are reading this, it means you already downloaded the VATK package using the following command:

git clone <https://github.com/centre-for-virus-research/VATK.git>

You can now navigate to the VirHosFilt folder:

cd VATK/VirHosFilt

Now you can use the commad pwd to output the string corresponding to the path:

pwd

The only thing you need to do in order to make the software functional on your home is to open the file VirHosFilt.py with you preferred text editor and change the second line:

installationFolder = "/home3/scc20x/Software/mySoftware/VATK/VirHosFilt"

so that the string produced by the pwd command replace that present between the double quote. You are now ready to go

**Running the GUI**

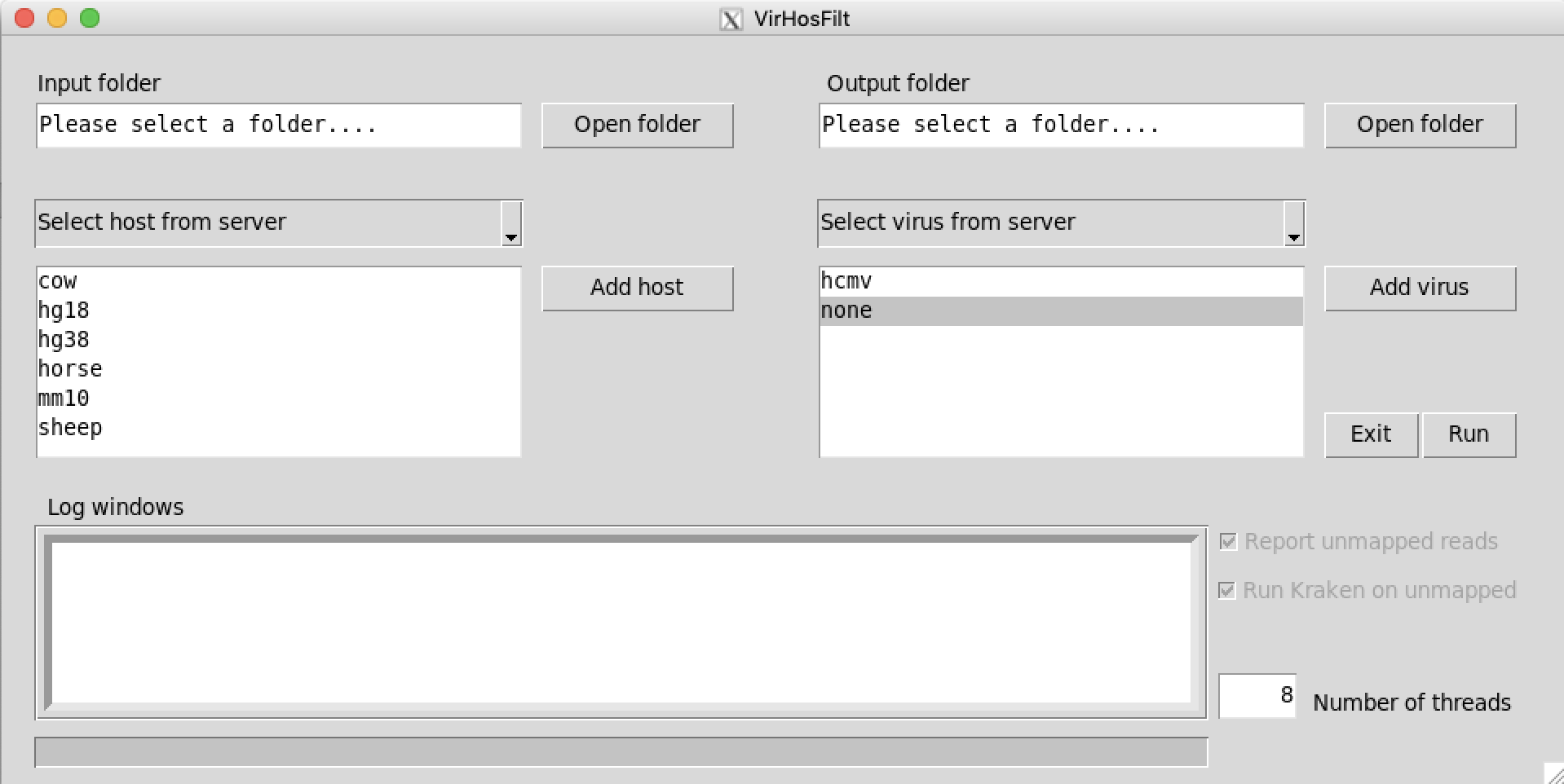
VirHosFilt comes with an easy to use graphic user interface (GUI). If you want to use the software on the server Alpha (this is the only way to access the host and virus genomes that are preloaded) you have to remember to access the server using the flag -Y, i.e:

ssh -Y [myUsername@alpha.cvr.gla.ac.uk](mailto:myUsername@alpha.cvr.gla.ac.uk)

After accessing the VirHosFilt on alpha you can launch the program by typing the following:

python VirHosFilt.py

The following window will open:



The “Open folder” button on the top left side can be used to select where all the input files that need to be analysed are. Make sure to double click on the folder where the files are. VirHostFilt will recognise the files belonging to the same paired dataset, but files must be named with the suffix \_1.fastq and \_2.fastq

The “Open folder” button on the top right side allows the user to select the output folder.

There are two drop-down menus reporting the statements “Select host from server” and “Select virus from Server”. If these options are selected, then in the two boxes below them will report the available host and virus genomes on alpha. Alternatively, these can be changed in “Select host locally” and/or “Select virus locally” and, in this case, host and virus genomes that have been created on your local installation path of VirHosFilt will be reported. See below how to create new host or virus VirHosFilt compatible genomes.

Two checkboxes will allow to report the host free reads that did not map on the selected virus reference genome and make a kraken analysis on them (these are available for the user only if a virus is selected from the list).

Finally the user can select the number of threads to be used and see the steps of the elaboration reported in the log windows.

**Testing the software**

You can test VirHosFilt by opening the input folder ./testFiles/input and selecting as the outputFolder ./testFiles/output by using the two “Open folder” button on the top of the GUI. The input files consist in a paired end dataset named read\_1.fastq and read\_2.fastq. These files contains few reads from a clinical sample of HCMV. Such a sample contains many human reads that need to be removed. While leaving the hosts and virus genome folders pointing on alpha (“Select host from server” and “Select virus from server” on the two drop-down menu”), select hg38 as the host genome (human genome version 38) and hcmv as virus (human cytomegalovirus, reference strain merlin). Tick “Report unmapped reads” and “Run Kraken on unmapped” to perform the complete analysis. Click the “Run” button to start the analysis.

The analysis should last not more than a couple of minutes and all the output files will be in the output folder. More precisely you will find the following:

reads\_hostAlignment.bam: the alignment file of the original reads on the host genome

reads\_virusAlignment.bam: the alignment file for the purged reads on the virus genome

read\_noHost\_1.fastq / read\_noHost\_2.fastq: the paired end dataset containing the host free reads.

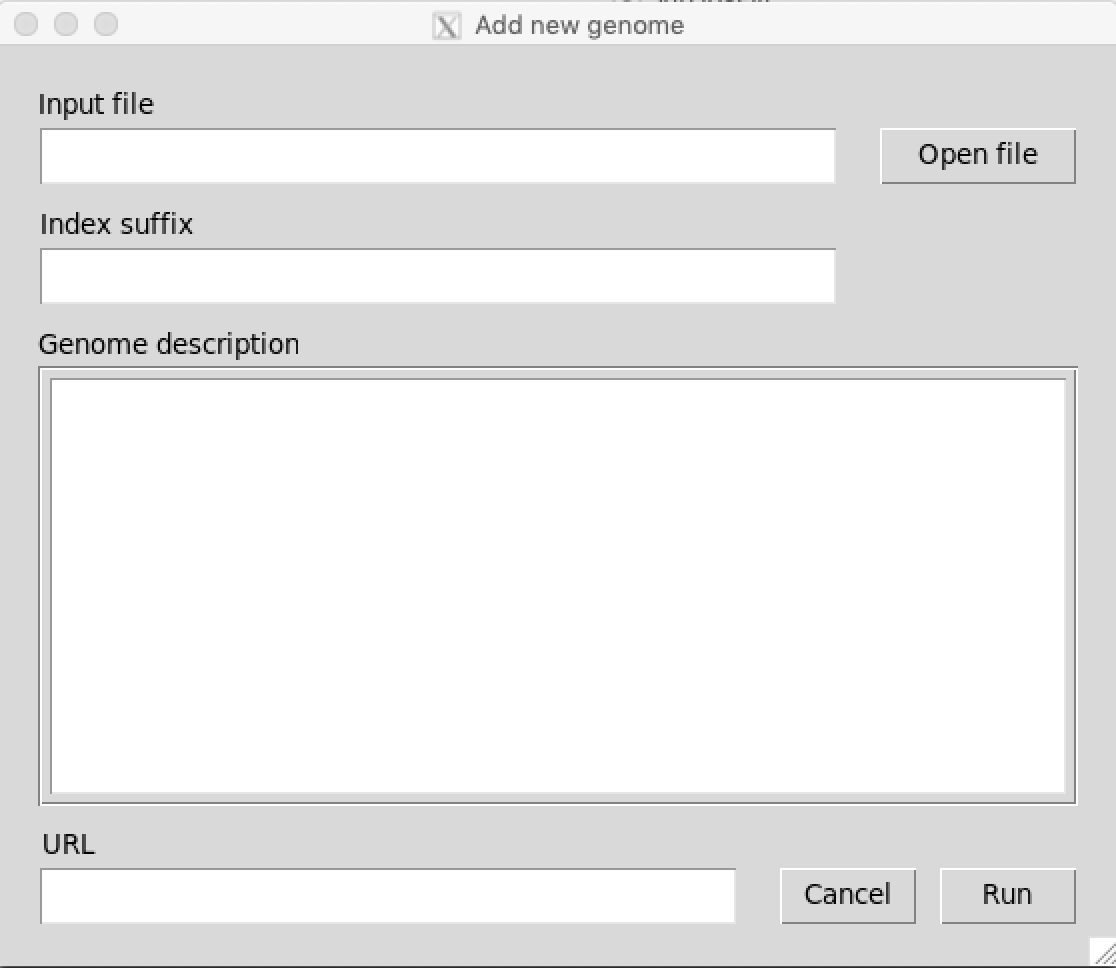
read\_unmapped\_1.fastq / read\_unmapped\_2.fastq: the reads from the read\_noHost\_1.fastq / read\_noHost\_2.fastq dataset that did not map on the reference virus genome

**Running your own sample**

Running your own sample is going to be as easy as running the test file. Just point to your input and output folder, select the host gneome and, if you want, the virus genome and tick or untick the additional analyses “Report unmapped reads” and “Run Kraken on unmapped” as desired. Then click Run to start the analysis. Please not that selecting a virus is not compulsory and you can just perform the analysis to obtain the host free fastq datasets.

**Creating a new host or virus VirHosFilt genome**

Close to the host and virus genome boxes there are the “Add host” and “Add virus” respectively. By clicking on them a new window will open that will look like the following:



You can use these windows to add (in your local VirHosFilt) a new host or virus genome. In order to add it the only info you need to provide is the fasta formatted file with the genome (you can navigate to it using the “Open file” button), the index suffix (a small, one word, name that you want to use for your host or virus that will appear in the relative box), a short description of the genome (where the genome comes from, why it is different by other present in the host or virus folder, if it is a specific strain and so on), the url where you downloaded the virus or host genome.

After providing all these information just press Run and wait for the elaboration to complete. Please be aware that for big host genomes the elaboration may take several hours.

Especially for the host genomes, it would probably be advisable to contact the Software administrator to have it added to the main server host folder so that this will be available for all the software users.