VireTap

A pipeline for general viral transcriptome detection in human RNA-seq data

A 03-713 Bioinformatics Practicum project. It is focused on virus detections of a given diseased model using RNA-seq data, integrating various bioinformatics tools including TopHat, Trinity, and BLAST.

Main contributors (A-Z last name): Yi-Yuan Lee, Chengyang Nie, Chengze Shen, Ayushi Sood, Shubhakar Tipireddy.

WARNING: This program needs to run on computers that have `slurm` activated (i.e. on PSC server, clustering servers).

Prerequisites

```
sra-toolkit/2.8.1 or above
java
PrgEnv-gnu/7.1
samtools/1.3 or above
pigz
oracle-jdk
gcc/5.3.0 or above
perl/5.18.4-threads
cufflinks/2.2.1
tophat/2.1.0 or above
bowtie2/2.2.7
jellyfish2/2.2.6
salmon/0.9.1
blat/v35
trinity/2.8.4
blast/2.6.0 or above
```

In addition, VireTap will come with BBMap-38.38 by brian-jgi and shc-3.8.7 by Francisco J. Rosales.

Download

Download the latest release to your local folder.

(Optional) Using shell script to download:

```
$ wget "https://github.com/c5shen/VireTap/releases/latest"
wget "https://github.com$(egrep 'archive.*tar\.gz' latest | cut -d '"' -f 2)"
rm latest
```

Installation

To install, extract files from the downloads.

```
$ tar -xvf [download].tar.gz
```

Then, `cd` into the newly made directory and `make` the binary executable.

```
$ make
```

This will output a binary executable `viretap` (if not provided) to the directory, and ask you for installation.

After Installation

You can choose to install the package to local, so you can execute VireTap as a standalone program anywhere in the system. This may need sudo permission (ask admin for more information).

Run VireTap

To run the program, execute the binary with the data access number you desire to perform viral transcriptome detection (for now, we only support human cell RNA-seq data).

```
$ ./viretap [ACCESSION]
```

Where [ACCESSION] refers to the accession number from NCBI for the particular RNA-seq dataset you are using.

Example 1

```
$ ./viretap SRR5787177
```

Check additional information by having -h or --help as the parameter.

```
$ ./viretap -h
```

Some modules can be modified by specifying them in parameters. For now we support the modification below:

```
    -i|--index <string>     Specify index folder for Tophat.
    -a|--accession <string> Specify the accession number.
    --num-cores <int>       Specify number of cores to use on node.
    --mem-trinity <int>     Number of GBs memory to use for Trinity.
```

Example 2

```
$ ./viretap -i /some/folder/index -a SRR5787177 --num-cores 10 --mem-trinity 80
```

Output

VireTap will download the GRCh38 homo sapien cdna index files from shared google drive, as well as a GI list of viruses for blast search. If you don't want to use the index files provided above, please specify your own index folder in the parameter.

VireTap will run tophat, Trinity, and blastn in sequence to find viral transcriptome in provided RNA-seq data. Then, it will construct a folder named [ACCESSION]_data, where all intermediate files are stored. A final blast output named [ACCESSION]_blast_output.txt will also be in that folder.

In addition, there will be a file named [ACCESSION]_blast_overview.txt, in which overviews of all viral hits are listed (no concentration/percentage of original RNA-seq analysis yet).

After each whole iteration of a job (for one dataset), there will be a slurm-[job].out report, which can be safely deleted.