Finding Viral Sequences in Pancreatic Tumor RNA Sequencing Data

Christopher Sontag

12/11/2022

Manual

Purpose

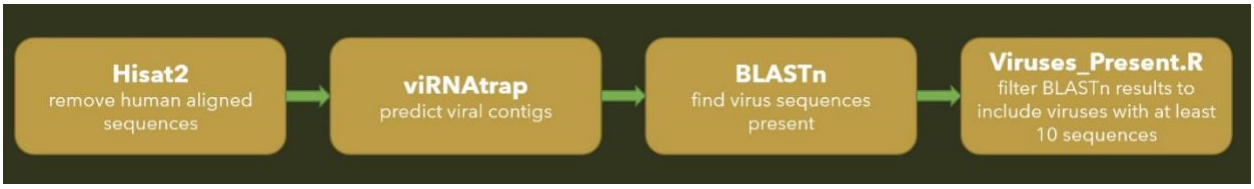
This guide is meant to provide directions to find and identify viral sequences in RNA sequencing data. In this guide, we are specifically focusing on finding and identifying viral transcript sequences in a pancreatic cancer tumor RNA sequencing dataset.

Presentation: Finding Viral Sequences in Pancreatic Tumor RNA Sequencing.pptx
Video Tutorial: Christopher_Sontag_Bioinformatics_Project_Video.mp4

Background

Cancer is an extremely common and often detrimental disease that can be difficult to treat. There are several mainstream therapies to address cancer, however the process of treating cancer is generally individualized as the characteristics of cancer differ greatly among individuals. Stimulating the immune system can better allow the body to find and destroy cancerous cells. Oncolytic Virus Therapy is a new form of cancer therapy that uses engineered viruses to remove problematic oncogenes and insert immune-stimulating genes into cancer cells (https://www.cancerresearch.org/treatment-types/oncolytic-virus-therapy). However, if the oncolytic virus is too similar to a virus the patient has recently had, the patient's immune system could potentially destroy these "good" viruses. In order to combat this, it is useful to know which viruses the patient has had so scientists can engineer the oncolytic viruses accordingly. When infected with a virus, the virus expresses its genes via RNA, which leads to viral RNA present within the host's body. During RNA (transcriptome) sequencing of cells from a host, resident viral RNA becomes sequenced along with the rest of the host's RNA. Bioinformatic tools can be used to detect viral RNA sequences (reads) from a host transcriptomic dataset. One strategy is to exclude host reads from a sequencing dataset, generate predicted viral contigs from the non-host reads, and determine what viruses the contigs came from (Elbasir *et al.* 2022).

Bioinformatic Solution



Our first step in finding viral sequences in human RNA-Seq data is to use Hisat2 to align sequences from the raw FASTQ file to a human reference genome. To more rapidly identify foreign RNA sequences that will only be a small proportion of all the reads in the dataset, we will remove all of the reads that map back to the human genome. One of the features of Hisat2 is that it can produce a file containing the reads that don't map back to the reference genome using the "un-conc unmapp" argument. The files used for the reference human genome were downloaded from (https://daehwankimlab.github.io/hisat2/download/#h-sapiens). Once human sequences are removed from the RNA-seq dataset, we use this much smaller dataset to efficiently look for viral sequences. Since we are only interested in the viral sequences present, it is important to remove the remaining bacterial RNA sequences. The viRNAtrap tool uses a deep-learning algorithm to predict viral reads from short read sequencing datasets. Its functionality allows prediction of viral reads without aligning them to a reference genome, allowing viral reads to be detected from all types of viruses.

The final step in this pipeline is to run BLAST to identify which species of viruses the previously predicted viral reads come from. Specifically, blastn will be used with our RNA sequencing data since its alignment works with nucleotide sequences. The blast web-app only works with small numbers of sequences, so BLAST Command Line was used. Blastn works by aligning nucleotide sequences to a chosen database of reference sequences. In this case we are interested in viral sequences so we will use the "ref_viruses_rep_genomes" database. This database was premade by NCBI and downloaded from (https://ftp.ncbi.nlm.nih.gov/blast/db/).

Dataset

In this example I am using human RNA-Seq data (Illumina short reads) derived from the tumor of a patient with pancreatic cancer (accession: SRR22062954). This dataset has nearly 20 million reads, and we expect it will contain transcript sequences from humans as well as different microbes in the host. I chose this dataset as it was readily available from SRA, however any human RNA-Seq data would work with this pipeline. The raw FASTQ file for this dataset was downloaded from (https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR22062954&display=download).

Installation of tools

1. Download the Anaconda command-line installer here. It is used to install the following command line tools on Mac operating systems.
2. Open the downloaded file and follow the directions to install Anaconda on your machine.
3. Install Hisat2 using the command:

```
conda install -c bioconda hisat2
```

4. Install viRNA trap using the following commands. This assumes that you have installed python and pip on your Mac.

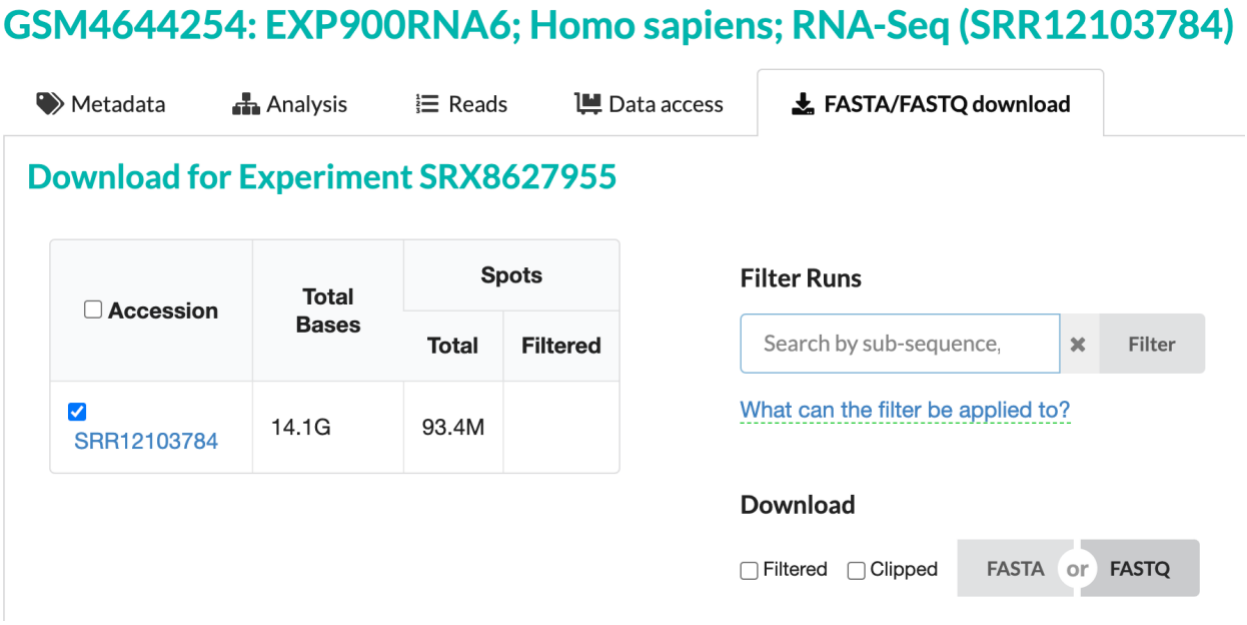
```
git clone https://github.com/AuslanderLab/virnatrap.git
cd virnatrap
conda create --name virnatrap python=3.7 pip
conda activate virnatrap
pip install .
```

5. Install BLAST command line using the command:

```
conda install -c bioconda blast
```

Running Pipeline

1. Download RNA Sequencing FASTQ file(s) from SRA by clicking on the FASTQ button under “Download”. You can download the data used in this guide here.



2. Download Hisat2 reference splice-aware genome from here by clicking on the “H. sapiens” link under “Index”.

- Index
  - H. sapiens
  - M. musculus
  - R. norvegicus
  - D. melanogaster
  - C. elegans
  - S. cerevisiae

3. Extract the contents of the reference genome (grch38_genome.tar.gz) into a new directory using the “gunzip” command.

```
gunzip grch38_genome.tar.gz
```

4. Run the make_grch38.sh script that is found in the previously extracted folder “grch38_genome” to complete the installation of the reference genome. This script is a wrapper that builds the genome into a format that is usable by Hisat2. It is also important to note that this wrapper script creates the genome index under the name “genome”.

```
./make_grch38.sh
```

5. To output all reads that remain unmapped against the human reference genome, make sure you are in the same directory that make_grch38.sh was run in, and run the “hisat2”

command as follows. The -U flag is used for unpaired FASTQ files, while --un flag is used to indicate a file to which all of the reads that don't map back to the reference genome, indicated with -x.

```
hisat2 -x genome -U ~/Bioinformatics_Project/SRR22062954.fastq --un unmapp.fastq
```

6. Once Hisat2 has finished running, move the resulting file to the input_fastq folder in the "virnatrap" directory that was created when viRNAtrap was installed.
7. viRNAtrap needs to be activated through Anaconda to use its functionality. Activate it using the command below

```
conda activate virnatrap
```

8. Run the viRNAtrap using the "virnatrap-predict" command as follows.

```
virnatrap-predict --input input_fastq/ --output output_contigs/
```

9. Download the "ref_viruses_rep_genomes" BLAST database using the update_blastdb.pl script (included when blast command line is installed) by running the command as follows.

```
update_blastdb.pl --decompress ref_viruses_rep_genomes
```

10. Run BLAST using the "blastn" command in the terminal as follows. With the -db flag set to ref_viruses_rep_genomes, the -query flag set to the path to the viRNAtrap output, and the -out flag set to "blast_output.csv".

```
blastn -db ref_viruses_rep_genomes -query ~/Bioinformatics_Project/virnatrap/
output_contigs/unmapp.txt -out blast_output.csv
```

11. The BLAST results are now found in the file "blast_output.csv". Run the Viruses_Present.R script that I wrote in R Studio to determine which virus species had at least 10 reads present in the raw FASTQ downloaded from SRA. The output file should look like the following.

Virus Species	Reads Present
>NC_032111.1 BeAn 58058 virus, complete genome	607
>NC_055235.1 Baboon cytomegalovirus OCOM4-37, complete genome, *** SEQUENCING	36
>NC_008168.1 Choristoneura fumiferana granulovirus, complete genome	33
>NC_043329.1 Diolcogaster facetosa bracovirus segment 29, complete sequence	20
>NC_022518.1 Human endogenous retrovirus K113 complete genome	16

Results

For the BLAST alignment, the default parameters were used meaning that at least 11 nucleotides needed to match a sequence in the database used to count in the "Reads Present"

column. 607 reads mapped to BeAn 58058 virus which is a member of the Poxyviridae family, along with smallpox and monkeypox. 36 reads mapped to baboon cytomegalovirus OCOM4-37, which is a member of the Herpesviridae family that also includes Epstein-Barr and human cytomegalovirus. This is interesting as Epstein-Barr has been shown to be correlated with some cancers (Bakkalci *et al.* 2020). 33 reads mapped to Choristoneura fumiferana granulovirus, which is a member of the Baculoviridae family. This family can make their way into mammal cells but are not known to be capable of replication in the cells. This was potentially caused by contamination during the library prep as a significant number of reads would not be produced unless the virus was able to replicate within the host. 20 reads mapped to Diolcogaster facetosa bracovirus, which is a member of the Baculoviridae family. A closely related virus, AcMNPV was cited as a potential vector for gene delivery applications (Matilainen *et al.* 2005). This may be useful for scientists to know as they will want to steer away from using the AcMNPV virus as a vector in any potential therapies. 16 reads mapped to human endogenous retrovirus K113, which is a member of the Retroviridae family. HIV and human T-lymphotropic virus are also members of Retroviridae. Interestingly, the closely related exogenous mouse mammary tumor virus causes breast cancer in mice (Xue *et al.* 2020), raising the possibility that the K113 virus may contain similar features in humans.

References

Bakkalci, D., Jia, Y., Winter, J. R., Lewis, J. E., Taylor, G. S., & Stagg, H. R. (2020). Risk factors for Epstein Barr virus-associated cancers: a systematic review, critical appraisal, and mapping of the epidemiological evidence. Journal of global health, 10(1), 010405. https://doi.org/10.7189/jogh.10.010405

Elbasir, A., Ye, Y., Schäffer, D. E., Hao, X., Wickramasinghe, J., Lieberman, P. M., … Auslander, N. (2022). Characterizing the landscape of viral expression in cancer by deep learning. BioRxiv. doi:10.1101/2022.06.26.497658

Matilainen, H., Rinne, J., Gilbert, L., Marjomäki, V., Reunanen, H., & Oker-Blom, C. (2005). Baculovirus entry into human hepatoma cells. Journal of virology, 79(24), 15452–15459. https://doi.org/10.1128/JVI.79.24.15452-15459.2005

Xue, B., Sechi, L. A., & Kelvin, D. J. (2020). Human Endogenous Retrovirus K (HML-2) in Health and Disease. Frontiers in microbiology, 11, 1690. https://doi.org/10.3389/fmicb.2020.01690