VireTap: Viral Detection in Human Disease Transcriptomes

Final Report- Bioinformatics Data Practicum

Team 5: Yi-Yuan Lee, Chengyang Nie, Chengze Shen, Ayushi Sood, Shubhakar Tipireddy

February 25, 2019

Contents

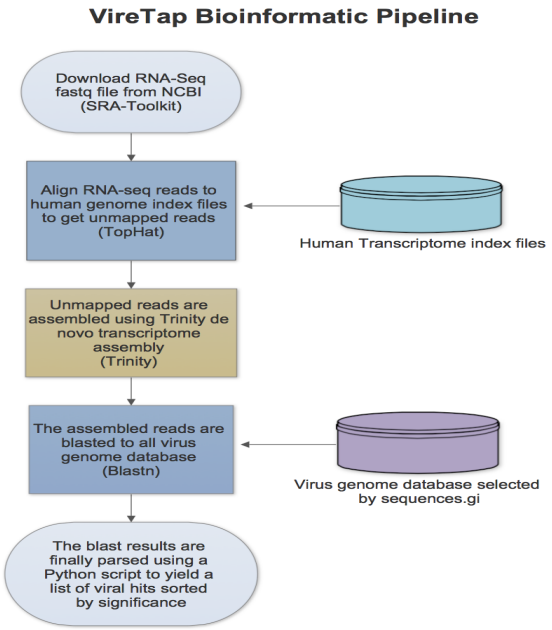
1 Introduction

Since the discovery of the Epstein-Barr Virus (EBV) as a causative agent for Burkitt's lymphoma in the 1960s [1], there have been several discoveries of viruses as the causative agents of many cancers as well as other diseases erstwhile classified as "non-infectious" [2]. The discovery of novel viral agents can be a huge step forward in understanding the pathology of a disease which can lead

to new drug development targets. While there are many bioinformatics tools to align sequence reads and assemble transcriptomes, to our knowledge there is no existing integrated pipeline to detect viral expression in a given disease RNASeq data.

VireTap is a simple pipeline which uses existing bioinformatics tools to analyze a given human disease transcriptome sample for viral expression patterns. On providing the SRA accession number for an RNASeq dataset, VireTap can download the data, run analyses using installed bioinformatics program (details in 2.2) and give the user a sorted list of viruses which are significantly aligned to the RNASeq data. Therefore, VireTap can be used as a starting point for exploratory analysis of a particular disease transcriptome and the results can be further quantified using other bioinformatics or experimental techniques.

The workflow of VireTap is summarized below. On getting the accession number, VireTap first downloads the RNASeq FASTQ file(s) using SRAToolkit's `fastq-dump` utility [3]. Then, human transcriptome index files are downloaded and the reads are aligned to these- the aligned reads are thrown away (as reads of human origin are not needed) and the unaligned reads are then assembled using Trinity de novo transcriptome assembly [4]. VireTap also downloads a `sequences.gi` file, which contains the NCBI accession numbers for all viral genomes taken from the NCBI Nucleotide database [5]. The assembled reads are then BLASTed against all viral genomes using `sequences.gi`. The BLAST output is then parsed using a Python script to yield a list of viral hits sorted by significance. The parsed output can be used by the user to direct further downstream analysis.



2 Methods and Materials

2.1 Datasets

The following datasets are for cancers with known viral causes. Both datasets are of human origin.

1. Kaposi's sarcoma: SRR5787177
2. Hepatocellular carcinoma: SRR4002942

2.2 Software Packages

2.2.1 SRA Toolkit

The SRA Toolkit and the source-code SRA System Development Kit (SDK) from NCBI allow us to download the data from NCBI and convert the sequencing data from SRA format into FASTQ or other formats. The most frequently used tool in SRA Toolkit is `fastq – dump` which can help user to download the sequencing data from NCBI in FASTQ format. In our pipeline, we also used `fastq – dump` to download raw data samples from NCBI.

For more details, please refer to:
https://ncbi.github.io/sra-tools/
Download and set the environment of SRA Toolkit:
https://github.com/ncbi/sra-tools

2.2.2 TopHat

TopHat is a bioinformatic sequence alignment tool for RNA-seq. It can align the RNA-Seq reads to mammalian-sized genomes using Bowtie and then analyze the mapping results. In our pipeline, we used TopHat to map the RNA-Seq data to human transcriptome index file to get unmapped reads for transcriptome assembly.

For more details, please refer to:
http://ccb.jhu.edu/software/tophat/manual.shtml
Download and set the environment of TopHat:
http://ccb.jhu.edu/software/tophat/tutorial.shtml

2.2.3 Trinity

Trinity is a bioinformatic tool for de novo reconstructing of transcriptomes from RNA-seq data. The tool depends on three independent modules: Inchworm, Chrysalis, and Butterfly, which can partition the sequence data into many de Bruijn graphs and then extract the full-length splicing isoforms from the graph. In our pipeline, Trinity is utilized to reconstruct the potential virus transcriptome based on the unmapped RNA-Seq reads.

For more details, please refer to:
https://github.com/trinityrnaseq/trinityrnaseq/wiki
Download and set the environment of Trinity:
https://github.com/trinityrnaseq/trinityrnaseq/releases

2.2.4 BLAST

Basic Local Alignment Search Tool (BLAST), developed by NCBI, identifies the species by taking nucleotide or protein sequence as query, and mapping to genome sequence. VireTap takes output file from Trinity and virus list to do BLASTN, which takes nucleotide as query. In the end, VireTap will produce brief report in overview file and detailed results in output file.

For more details, please refer to:
https://blast.ncbi.nlm.nih.gov/Blast.cgi
Download and set the environment of BLAST:
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

3 Using the Pipeline

3.1 Dependencies

The following programs should be installed and added to path before running the pipeline:

```
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
```

3.2 Downloading and Installation

VireTap can be downloaded directly from GitHub. Optionally, it can also be downloaded from the shell:

```
$ wget "https://github.com/c5shen/VireTap/releases/latest"
```

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

Extraction of files from the download:

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

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

3.3 Running VireTap

The binary is executed with one step, and only requires the SRR accession number of the data to be analyzed:

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

There is a command line help option:

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

The following command line modifications are supported:

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

3.4 Output

VireTap will download the **GRCh38 homo sapien cdna** index files from the group's Google Drive, as well as a GI list of viruses for blast search. The user can also specify their own TopHat index files as specified in 3.3.

The main outputs of VireTap are the following files:

1. `[ACCESSION]_blast_output.txt`: BLAST results for the assembled reads, containing all data provided by BLAST about the matches.
2. `[ACCESSION]_blast_overview.txt`: A list of the viral BLAST hits sorted by significance.

4 Results

The top 15 hits in the BLAST overview for Kaposi's sarcoma were:

ID	NAME	SCORE	EVALUE
KT271460.1	Human herpesvirus 8 strain ZM114, partial genome	4927	0
KT271468.1	Human herpesvirus 8 strain ZM130, partial genome	4927	0
KT271458.1	Human herpesvirus 8 strain ZM106, partial genome	4927	0
KT271464.1	Human herpesvirus 8 strain ZM121, partial genome	4927	0
KT271457.1	Human herpesvirus 8 strain ZM102, partial genome	4927	0
KT271454.1	Human herpesvirus 8 strain ZM027, partial genome	4922	0
KT271459.1	Human herpesvirus 8 strain ZM108, partial genome	4922	0
KT271462.1	Human herpesvirus 8 strain ZM117, partial genome	4922	0
KT271455.1	Human herpesvirus 8 strain ZM091, partial genome	4922	0
KT271466.1	Human herpesvirus 8 strain ZM124, partial genome	4872	0
KT271453.1	Human herpesvirus 8 strain ZM004, partial genome	4872	0
KT271461.1	Human herpesvirus 8 strain ZM116, partial genome	4867	0
KT271463.1	Human herpesvirus 8 strain ZM118, partial genome	4867	0
U93872.2	Kaposi's sarcoma-associated herpesvirus glycoprotein M,...	4855	0
KF588566.1	Human herpesvirus 8 clone BrK.219#1 4, complete genome	4850	0

The complete overview file can be found on GitHub.

The top 15 hits in the BLAST overview for Hepatocellular carcinoma were:

ID	NAME	SCORE	EVALUE
JX661491.1	Hepatitis B virus isolate SH1222-C10, complete genome	3642	0
KM213037.1	Hepatitis B virus isolate X15-k1, complete genome	3626	0
JX661490.1	Hepatitis B virus isolate SH1217-C8, complete genome	3615	0
JX661487.1	Hepatitis B virus isolate SH1207-C3, complete genome	3587	0
JX661489.1	Hepatitis B virus isolate SH1218-C9, complete genome	3576	0
AB206816.2	Hepatitis B virus DNA, complete genome, strain: Hiros...	3570	0
JX661488.1	Hepatitis B virus isolate SH1212-C5, complete genome	3570	0
JX661492.1	Hepatitis B virus isolate SH1215-C7, complete genome	3565	0
JX661496.1	Hepatitis B virus isolate SH1208-C4, complete genome	3565	0
JX661493.1	Hepatitis B virus isolate SH1226-C13, complete genome	3559	0
JX661495.1	Hepatitis B virus isolate SH1223-C11, complete genome	3554	0
JX661486.1	Hepatitis B virus isolate SH1225-C12, complete genome	3548	0
AB206817.2	Hepatitis B virus DNA, complete genome, strain: Toran...	3542	0
AF384371.1	Hepatitis B virus isolate G376-7, complete genome	3537	0

The complete overview file can be found on GitHub.

5 Discussion

VireTap detects Hepatitis B expression in hepatocellular carcinoma and HHV8 in Kaposi's sarcoma, which is in line with empirical knowledge about these diseases [6]. Hence, these datasets act as a positive control for the pipeline and show that if there is a major viral cause behind a disease, then VireTap can find the corresponding viral species in transcriptome datasets for that disease.

A weak point of VireTap is its extremely high sensitivity; in other words, while the top results are for the viruses which actually cause the disease, there are a lot of hits for unrelated viruses such as Zika, Dickey phage, and viruses of plant and animal origin which could not have been actually present in the human sample. It is possible that these are just matches in small regions which are conserved across viruses; in any case, VireTap would greatly benefit from a filtering mechanism for such BLAST hits which are not likely to have any significance. The threshold for significance should also be ideally modifiable by the end user, depending on the disease being analysed and the specifics of the

transcriptome dataset.

While VireTap is a good first step for finding viral expression in disease transcriptomes, there is currently no way to quantify the expression of viral sequences, or to check the coverage of viral matches with the original transcriptome. Adding these functionalities to VireTap is a natural next step for the project and will make it more scientifically rigorous and useful.

Computationally speaking, we could improve VireTap by making it more accessible for all platforms and having more modularity depending on the configuration of the user's system. For instance, the installation steps could be different depending on the user's system, `slurm` availablity, etc. Speed-wise, the bottleneck in the current implementation is TopHat as aligning the reads to the entire human transcriptome takes a lot of time; we will also look into optimizing this.

References

[1] Michael Anthony Epstein, Bert G Achong, and Yvonne M Barr. "Virus particles in cultured lymphoblasts from Burkitt's lymphoma". In: *The Lancet* 283.7335 (1964), pp. 702–703.

[2] André J Nahmias and Richard J O'Reilly. *Immunology of Human Infection: Part II: Viruses and Parasites; Immunodiagnosis and Prevention of Infectious Diseases*. Vol. 9. Springer Science & Business Media, 2012.

[3] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, et al. "The sequence read archive". In: *Nucleic Acids Research* 39.suppl_1 (2010), pp. D19–D21.

[4] Brian J Haas, Alexie Papanicolaou, Moran Yassour, et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis". In: *Nature Protocols* 8.8 (2013), p. 1494.

[5] "Database resources of the national center for biotechnology information". In: *Nucleic Acids Research* 46.D1 (2017), pp. D8–D13.

[6] Véronique Bouvard, Robert Baan, Kurt Straif, et al. "A review of human carcinogens—Part B: biological agents". In: *The Lancet Oncology* 10.4 (2009), pp. 321–322.