**Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution**

**User Manual**

Section I. Preliminary preparation

Section II. Suggested working procedures

Section III. Output file format

Authors: Daniel WH Ho & Irene OL Ng

Department of Pathology and State Key Laboratory for Liver Research, The University of Hong Kong

Contact: dwhho@hku.hk

Last update: 20 January 2015

**Section I. Preliminary preparation**

**a) Downloading the human and virus reference genome**

User should get ready the human (human.fa) and virus (virus.fa) reference genome in FASTA format.

**b) Installation of BWA (http://bio-bwa.sourceforge.net/)**

User should choose the latest version of BWA (currently 0.7.10). Following the instructions from the provider. BWA-MEM will be used by Virus-Clip. Index should be built for the virus reference genome using the following command:

**bwa index virus.fa**

The index generated should be placed under the same directory as the virus reference genome.

**c) Installation of SAMTools (http://www.htslib.org/)**

User should choose the latest version of SAMTools (currently 1.1). Following the instructions from the provider.

**d) Installation of standalone BLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/)**

User should choose the latest version of standalone BLAST (currently 2.2.30). Following the instructions at http://www.ncbi.nlm.nih.gov/books/NBK52640/ for installation. BLASTN will be used by Virus-Clip. BLAST database should be built for the human reference genome using the following command:

**makeblastdb -in human.fa -dbtype nucl -out human**

The BLAST database generated should be placed under the same directory as the human reference genome.

**e) Installation of ANNOVAR (http://www.openbioinformatics.org/annovar/)**

User should choose the latest version of ANNOVAR (currently 2015Jan06). Following the instructions from the provider.

**f) Creating the home directory for Virus-Clip result**

User should create home directory (e.g. /virus\_clip) for Virus-Clip execution. Under this directory, sub-directories of align (e.g. /virus\_clip/align) and result (e.g. /virus\_clip/result) should be made. These 2 sub-directories will store the read alignment and viral integration detection result respectively.

**Section II. Suggested working procedures**

**Step 1.**

**Setting the path and parameter information in virus\_clip.sh**

*Users should amend the parameters in the shell script BEFORE the execution of the script.*

In the tools section of the script, user should fill in the path information for:

1. BWA executable

2. SAMTools executable

3. virus\_clip.pl (provided in Virus-Clip)

4. BLASTN executable

5. ANNOVAR annotate\_variation.pl

In the resources section of the script, user should fill in the path information for:

1. Virus reference genome (virus.fa)

2. BLAST database for human reference genome

3. ANNOVAR database for refGene

In the parameter section of the script, user should fill in the information for:

1. Sequence read file suffix (e.g. .fastq.gz)

**Step 2.**

**Executing virus\_clip.sh**

User can execute virus\_clip.sh through the following command:

**bash virus\_clip.sh path\_to\_seq\_data path\_to\_result xxx library\_flag**

where

path\_to\_seq\_data Path to directory storing sequence file

path\_to\_result Path to home directory for Virus-Clip result

xxx Name prefix of the sequence file

library\_flag 1 for single-end, 2 for paired-end

For single-end data, the sequence file should be named as xxx.fastq.gz. For paired-end data, the 2 sequence files should be named as xxx\_1.fastq.gz and xxx\_2.fastq.gz. Read alignment in SAM format (xxx.bwa.sam) will be created under the align sub-directory. Directory with name prefix of the sequence file will be created under the result sub-directory (e.g. /virus\_clip/result/xxx) and Virus-Clip output file (virus\_clip.out) will be generated, which stores the viral integration events detected.

**Section III. virus\_clip.out file format (tab-delimited text)**

Column 1 Left element of the integration event

Column 2 Chromosome for the left element

Column 3 Breakpoint position for the left element

Column 4 Sequence for the left element

Column 5 Right element of the integration event

Column 6 Chromosome for the right element

Column 7 Breakpoint position for the right element

Column 8 Sequence for the right element

Column 9 Supporting soft-clipped read count

Column 10 Affected human gene region

Column 11 Affected human gene