# 1. Install

Download all the files into one folder and run main.pl using "perl /absolute\_path\_of\_main.pl/main.pl parameter1 parameter2 parameter3 ......". The users should install the packages used in perl and python programs, such as PerlIO::gzip, Getopt::Long, File::Basename, etc.

# 2. One demo

A demo has been uploaded. Users can download the file "demo.rar" and unzip it. We have add an file named "used.cml" in each folder. used.cml contains the command lines used in that folder. Please note that users should replace the absolute path of all the files in each script to run the demo.

# 3. A Step-to-step protocol of the HIVID2 pipeline

## 3.1 The main program: main.pl

## 3.2 Parameters

**-o** output directory path  
**-l** a file containing sample\_id, library\_id and FC\_id  
**-stp** step number (1/2/3/4)  
**-c** parameter configuration file  
**-filter** whether to filter the repeated comparison reads. Here, only the repeated comparison reads on the human genome are filtered. The repeated comparison reads on the HBV genome are not filtered. However, in the result, the reads of repeated alignments on the HBV genome will be discarded, and the only aligned reads on the corresponding human genome will be retained.  
**-f** this parameter is currently useless，please do not use it.

## 3.3 Description of several predefinding files

### (1) -C the Configure file

This configure file difined the referece genomes and alignment parameters used in step3. The users can make their own configure file. But we have involved some configure files which is named as Config\* in the same folder of main.pl. Below is the description of the configuration file:  
soap: the path of the soap2 program  
ref\_virus: the path of soap2 index of virus reference genome  
ref\_human: the path of soap2 index of human reference genome  
insert\_sd: the standard deviation of the insert size for the sequencing library  
virus\_config: the parameters of soap2 corresponding to different read length; for example, "150;150:-l 50 -v 5 -r 1" means when the read length is 150 bps, then soap2 will use the parameter "-l 50 -v 5 -r 1"; please note that read length is set at sample.list under the folder step1.

### (2) -l a file containing sample\_id, library\_id and FC\_id

It can be named as any name and simply write as the sample name in three column. For example, a file named "list" and contain a line with three columns: SRR12345 SRR12345 SRR12345

## 3.4 Step to step tutorial

### 1st step

Manually create a file named "list" in the output directory. Then manually create a folder named step1, and create a file named sample.list in folder step1. The location of sample.list should be step1/sample.list. Note that the path in the sample.list should be absolute full path and the word in the first four columns should be the same as that in the file "list". Please note that step1 is all done by hand and do not use main.pl in this step. Below is an example of sample.list:

Sample FC Lane Libray read\_length library\_size  
SRR12345 SRR12345 SRR12345 SRR12345 110;110 170 /absolute\_path/bkread1.fq.gz /absolute\_path/bkread2.fq.gz

### 2nd step

perl /absolute\_path\_of\_main.pl/main.pl -o -l list -stp 2 -c configuration\_file

Example: perl main.pl -o test2 -l list -c ConfigHPV\_19 -stp 2

Note: 1) After running this command, the directory for each sample will be generated under the step2 folder. A shell script for each sample will be generated under corresponding folder. These shell scripts could be manually submitted onto to the server using qsub.

### 3rd step

perl /absolute\_path\_of\_main.pl/main.pl -o -l -stp 3 -c configuration\_file Example: perl main.pl -o test2 -l list -c ConfigHPV\_19 -stp 3

Note: (1) After running this command, a folder named after the sample name will be automatically generated in step3 folder. In each sample folder, there are three scripts Human\_virus\_soap.sh and station.sh. These scripts could be manually delivered to the sge system using qsub. The user should run Human\_virus\_soap.sh first before running station.sh.

(2) Run Dataproduction.sh manually to get the data production report including the data amount, data quality, sequence alignment profiles.

### 4th step

perl /absolute\_path\_of\_main.pl/main.pl -o -l -stp 4 -c configuration\_file -filter -fa1 -fa2 Example: perl main.pl -o test2 -l list -c ConfigHPV\_19 -stp 4 -filter -fa1 human.fa -fa2 hpv.fa

Note:

(1) -fa1 and -fa2 is compulsory. Once this step is finished, a folder named after the sample name will be generated in step4 folder. There is a shell script in each directory. These scripts could be executed by "sh xx.sh" or manually delivered to the SGE system using qsub.

(2) It should be noted that there are a file named "ref.list" in the same folder of main.pl. "ref.list" must contain all the ID of reference genomes used in the sequence alignment of step3 and step4, or the user will get error or uncompleted results in \*human\_bk.final.stp2.uniq2.final during the procedure of deep removing PCR-duplications. We have involved some predefined reference names in ref.list, but the users should add the references names used in their own experiments. In the ref.list, each ID should be followed by an underline, for example "chr1\_".

## 3.5 Result file and the format descript

The path of the files of final results:

The file of human breakpoint: step4/\*/human/ breakpoint/\*human\_bk.final.stp2.uniq2.final

The file of virus breakpoint: step4/\*/virus/breakpoint/\*HBV\_bk.final.stp2.uniq

Format description of the result file:

1st column is the number of the chromosome where the breakpoint located.

2nd column is Specific position coordinates

3rd column is the pair amount of left support reads

4th column is the pair amount of right support reads

5th column is the pair amount of discordant support reads

6th column is total number of all support reads

7th column is normalized pair amount of left support reads (normalized\_value =support\_reads\_number / effective\_reads\_number\_of\_the\_sample)

8th column is normalized pair amount of right support reads (logarithmic) normalized value

9th column is normalized pair amount of discordant support reads

10th column is total number of reads (logarithmic) normalized value

11th column is reads id of left support reads

12th column is reads id of right support reads

# 4. Advanced analysis

After obtaining the integration sites, HIVID2 allows the user to decide whether to automatically perform advanced analysis using the identified virus integrations.

(1) Manually seprate result folders of step4 into two groups, For example, tumor and normal, or other user-definednames. If you ran tumor and normal samples in a single run, then you may move each sample (each sample has a folder in step4) into the tumor or normal folder; if you iniatially ran tumor and normal samples seprately during step4, then you can simply use the step4 folder of tumor and normal of each run.

(2) Run advanced analysis #First， run Analyse.sh, generatint R scripts and the relevant files. sh /absolute\_path\_of\_main.pl/advanced\_analysis/Analyse.sh /absolute\_path/tumor /absolute\_path/normal  
#Second, run the generated R scripts Rscript xxx.R

Note: If you want to get the graph one by one, please separate the script and change parameters. You can also run it line by line, and modify the parameters by yourself.

# 5. Other tips

(1) In order to help the users to track the data processing, HIVID2 retained some intermediate procedure files during running of the pipeline. It may cause big hard disk consuming when deal with large amount of data such as WGS data. Fortunately, The users can can remove most of intermediate files of previous steps when running step4. When running step4, the user can remove all the files named "\*paired.gz" and "\*unpaired.gz" in step2, all the files named "\*soap.gz" in step2. After completing step4, all the files except the files of final results could be deleted. But before deleting, the users should make sure they don't need them later.

(2) About setting the length in sample.list: It is OK to set the length based on the raw reads, But it will be better set the reads length after removing the adapter. Actually, users can set the read length in sample.list after completing step2 because this value of length will not be used in step2 but used in step3. And in step2, adapters will be removed.

(3) There is a file named "tfbsConsSites.txt" in the advanced analysis. This file cannot be uploaded onto github due to the size limitation. But the user could download this file from Table browser of UCSC.

(4) HIVID2 works quite well for virus-capture sequencing data. For WGS data, sometimes the used memory might be too large. In this case, the users may need to separate the fastq data into several parts before input into HIVID2 for step1,step2 and step3; then the users can merge the data of step3 for the separated parts to run step4. For WGS data, the users could alternatively first remove human reads or HBV reads before running HIVID2.

(5) The programs of soap2 are also uploaded in this repository. The method of buliding index of soap2 is as following: 2bwt-builder xx.fa

# 6. Citation

Xi Zeng, Linghao Zhao, Chenhang Shen, Yi Zhou, Guoliang Li, Wing-Kin Sung, HIVID2: an accurate tool to detect virus integrations in the host genome, Bioinformatics, Volume 37, Issue 13, 1 July 2021, Pages 1821–1827, <https://doi.org/10.1093/bioinformatics/btab031>