Manual of VirusRecom

Detecting recombination of viral lineages using information theory

Home page: https://github.com/ZhijianZhou01/VirusRecom

Zhi-Jian Zhou, Ye Qiu, Chen-Hui Yang, Sheng-Bao Ye and Xing-Yi Ge

Version 1.0 || May 24, 2022

Content

1. Introduction

1.1. Background

Recombination is common in viruses. In general, recombination can be divided into two categories: homologous recombination including normal homologous recombination and abnormal homologous recombination, and non-homologous recombination. Homologous recombination occurs between homologous or similar sequences and plays a major role in recombination. The normal homologous recombination was defined as that recombination region is homologous or similar. However, aberrant homologous generally results in sequence deletion or duplication due to the recombinaton region is not in the homologous region.

Identifying homologous recombinations from highly similar sequences is a challenge, due to the uncertainty of an emerging genomic variance originating from recombination or *in-situ* mutation. Herein, we present VirusRecom, an efficient tool for recombination analysis of viral genome with high similarity based on information theory. The new method evaluates the likelihood of recombination by quantifing recombination contribution using weighted information content (WIC).

1.2. Functions

VirusRecom was written by Python 3, the releases were created by Pyinstaller. The functions are:

(I) Calculate the recombination contribution of each polymorphic sites.

(II) Identify the recombination events and potential recombination region with *p*-value.

(III) Scan the potential recombination breakpoint.

2. Download and install

VirusRecom and all the updated versions is freely available at https://github.com/ZhijianZhou01/VirusRecom. After obtaining the program, users could directly run the program in Windows, MacOS or Linux (Ubuntu 16.04 or more) systems without installation.

In general, the executable file of VirusRecom is located at the folder of Main. Then, just double click the virusrecom.exe (windows system) or virusrecom (Linux or MacOS system) to start. If you could not get permission to run virusrecom on Linux or MacOS system, you could change permissions by **chmod -R 777 virusrecom**.

3. Description of parameters

3.1. Getting help

VirusRecom is a command line interface program, users can get help documentation of the software by entering **virusrecom -h** or **virusrecom --help**.

```
usage:
virusrecom [-h] [-a ALIGNMENT] [-q QUERY] [-l LINEAGE] [-g GAP]
[-m METHOD] [-w WINDOW] [-s STEP] [-mr MAX_REGION]
[-p PERCENTAGE] [-b BREAKPOINT] [-bw BREAKWIN]
[-t THREAD] [-y Y_START]
```

Example of usage:

(i) If the input-sequence data was not aligned:

```
virusrecom -q XE.fasta -l Lineage_Dir -g n -m p -w 100 -s 20 -t 2
```

(ii) If the input-sequence data has been aligned:

```
virusrecom -a alignment.fasta -q XE_ -l lineage_name_list.txt -g n -m p -w 100 -s 20
```

3.2. Parameters

Parameters	Description
-h, --help	show this help message and exit.
-a	FilePath of an aligned sequence set (*.fasta format) containing all sequences used for analysis, then the alignment will be skipped. Default is null. If using, name of each sequence in aligned sequence set requires containing the mark (a unique string) of the lineage.
-q	FilePath of query lineage (potential recombinant, *.fasta format). Note, if the '*-a* alignment.fasta' has been used, please enter the mark (a unique string) of queried recombinant here, such as '*-q* XE_', not a FilePath. Using '-q auto' and all lineage will be scanned as potential recombinants in turn.
-l	DirPath of reference lineages. One sequence file (*.fasta format) per lineage, and each lineage could contain multiple sequences. Note, if the '*-a* alignment.fasta' has been used, please enter a text file containing the marks (a unique string) of lineages here, not a DirPath.
-g	Gaps (-) in the alignment were used in analysis? '*-g* y' means to reserve gaps, and '*-g* n' means to delete gaps.
-m	Scanning method of recombination analysis. '*-m* p': using polymorphic sites only, '*-m* a': using all the monomorphic sites and polymorphic sites.
-w	Number of nt sites per sliding window. Note: if the '*-m* p' has been used, *-w* refers to the number of polymorphic sites per windows.
-s	Step size for scanning these sites. Note: if the '*-m* p' has been used, *-w* refers to the number of polymorphic sites per jump.
-mr	The maximum allowed recombination region. Note: if the '*-m* p' method has been used, it refers the maximum number of polymorphic sites contained in a recombinant region.

Parameters	Description
-cp	The cutoff threshold of proportion (*cp*, default was 0.9) for searching recombination regions when mWIC/EIC >= *cp*, the maximum value of *cp* is 1. For detection in genus level, about 0.5 is recommended.
-b	Whether to run the breakpoint scan of recombination. '*-b* y': yes, '*-b* n': no. Note: this option only takes effect when '*-m* p' has been specified!
-bw	The window size (polymorphic sites, default is 200) used forbreakpoint scan. The step size is fixed at 1. Note: this option only takes effect when '*-m* p *-b* y' has been specified!
-t	Number of threads used for the multiple sequence alignments (MSA), default is 1.
-y	Specify the starting value of the Y axis in the picture, the default is 0.

4. Example of usage

4.1. Unaligned input-sequences

VirusRecom owns a pipeline built in to handle unaligned sequences. In this case, multiple sequence alignment is performed by MAFFT (Katoh & Standley, 2013) with alignment strategy of "auto". All the unaligned sequences from the query lineage need to contained in a file (*.fasta format), such as XE.fasta. For other lineages as reference lineages, the unaligned sequences from each lineage also need to contained in a file, and all the sequence files are placed in the same folder, for example (D:\Test\ lineages).

```
AY.45.fasta
B.1.617.2.fasta
B.1.1.7.fasta
BA.1.fasta
BA.2.fasta
```

Then, execute the command to detect recombinaton events, for example.

```
virusrecom -q D: \Test\XE.fasta -l D:\Test\ lineages -g n -m p -w 100 -s 20 -t 2
```

4.2. Aligned input-sequences

Users can also provide an independent aligned sequences dataset which was performed from any other alignment program, including all the sequences from query lineage and reference lineages. Notably, each sequence in the aligned sequences dataset needs to contain a unique mark representing its lineage, for example (XE-others.fasta).

```
>XE_hCoV-19/England/PHEC-YYR4GXD/2022
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>XE_hCoV-19/England/LSPA-3C834E6/2022
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>AY.45_hCoV-19/SouthAfrica/NICD-N18409/2021
-----------taaacgaactttaaaatctntgnggctgtcactcggctgcatgcttagtgcactcacgcag
>AY.45_hCoV-19/SouthAfrica/NICD-N18483/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>AY.45_hCoV-19/Denmark/DCGC-173261/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>B.1.617.2_hCoV-19/India/KA-CRL-KIMS-245/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>B.1.617.2_hCoV-19/Turkey/HSGM-F14312/2022
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>B.1.1.7_hCoV-19/Turkey/HSGM-FS7372/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgtatgcttagtgcactcacgcag
>B.1.1.7_hCoV-19/USA/FL-BPHL-6009/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>BA.1_hCoV-19/USA/MN-CDC-IBX539664392073/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>BA.1_hCoV-19/India/KA-SEQ_8548_S89_R1_001/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>BA.1_hCoV-19/Mexico/CHH_InDRE_FB2199_E08113577401_S11007/2022
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>BA.2_hCoV-19/Denmark/DCGC-294641/2021
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
>BA.2_hCoV-19/Malaysia/MGVI_GS0822/2022
gatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcag
```

Then, prepare a text document that labels all the names of reference lineages, for example (lineages.txt).

```
AY.45_
B.1.617.2_
B.1.1.7_
BA.1_
BA.2_
```

It is recommended to include the character "-" in the mark, which ensures that the mark of each lineage is unique. Notably, the mark of query lineage cannot be placed in the lineages.txt.

Then, execute the command to detect recombination events, for example.

```
virusrecom -a XE-others.fasta -q XE_ -l D:\Test\lineages.txt -g n -m p -w 100 -s 20
```

Note, the mark of query lineage was specified in the command, and it is recommended to include the character "-" in the mark.

4.3. Output result

As the Unaligned input-sequences above as an example. The output directory is automatically created and is located under the input directory. There are three sub-directories and the aggregated report (Possible recombination event in XE_.txt) under the output directory, including run_record, WICs of sites and WICs of slide_window.

(I) In the directory of run_record, the alignment file created by MAFFT is reserved. If **-g** n is specified, and the Record of deleted gap sites_*.txt file containg all the gap sites will be created. Besides, If **-m** p is specified, and the Record of same sites in aligned sequence_*.txt file containg all the same sites will be created.

(II) In the directory of WICs of sites, the *_WIC contribution from lineage in

sites.pdf and the *_WIC contribution from lineage in sites.xlsx are used to record the WIC value for each site.

(III) In the directory of WICs of slide_window, the *_WIC contribution from lineage in sliding window.pdf and the *_WIC contribution from lineage in sliding window.xlsx are used to record the mean WIC of each sliding window.

(IV) The identified recombination events and region are aggregated in the *_Possible recombination event in XE_.txt file.

Besides, if **-b** y is specified, for eaxmple, **-b** y **-bw** 200, then VirusRecom will perform the search of recombination breakpoint. The negative logarithm of *p*-value in each site is in the *_-lg(p-value) for potential breakpoint.pdf and *_-lg(p-value) for potential breakpoint.xlsx.

5. A simple running example

https://github.com/ZhijianZhou01/VirusRecom/tree/main/eaxmple/recombination_test_data.zip

Take synthetic data above as an example, the compressed file contains two files, one is sequence dataset and the other is a text file contains the the names of different lineages. The sequence dataset is already aligned and contains ten lineages, and each lineage owns 100 sequence samples. Besides, the known recombinant lineage has the label of "query_recombinant" in sequence names, and each other lineage has its own label in sequence names.

Then, we begin our analysis of potential recombination lineages. We take the windows system as an example, and assume that the root directory of the sample data is J:\test.

Besides, assume that the path to the executable of virusrecom is in J:\app\VirusRecom\main\virusrecom.exe

(i) Open the Windows PowerShell, and execute the following command.

J:\app\VirusRecom\main\virusrecom.exe -a J:\test\recombination_test_data\lineages_data_alignment.fas -l J:\test\recombination_test_data\reference_lineages_name.txt -q query_recombinant -g n -m p -w 100 -s 20 -o J:\test\recombination_test_data\out

(ii) Then, virusrecom begins calculating the recombination contribution of each reference lineage to the query lineage.

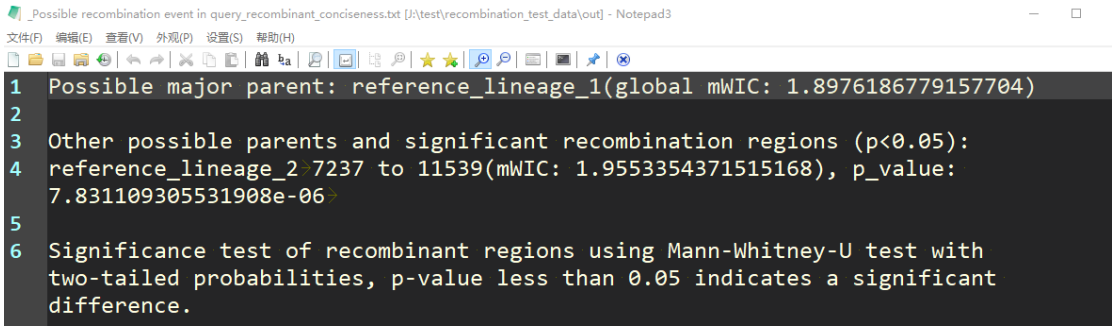
```
Windows PowerShell
-----------------------------------------------------
☆ Example of use ☆
  (1) If the input-sequence data was not aligned:
      virusrecom -q XE.fasta -l Lineage_Dir -g n -m p -w 100 -s 20 -t 2

  (2) If the input-sequence has been aligned:
      virusrecom -a alignment.fasta -q XE_ -l lineage_name_list.txt -g n -m p -w 100 -s 20
-----------------------------------------------------
PS C:\Users\j> J:\app\VirusRecom\main\virusrecom.exe -a J:\test\recombination_test_data\lineages_data_alignment.fa
s -l J:\test\recombination_test_data\reference_lineages_name.txt -q query_recombinant -g n -m p -w 100 -s 20 -o J:
\test\recombination_test_data\out
-----------------------------------------------------
  Name: Virus Recombination(VirusRecom)
  Description: Detecting recombination of viral lineages (or subtypes) using information theory.
  Version: 1.0 (2022-04-18)
  Author: Zhi-Jian Zhou
-----------------------------------------------------

Namespace(alignment='J:\\test\\recombination_test_data\\lineages_data_alignment.fas', breakpoint='n', breakwin=200
, gap='n', lineage='J:\\test\\recombination_test_data\\reference_lineages_name.txt', max_region=1000, method='p',
outdir='J:\\test\\recombination_test_data\\out', percentage=0.9, query='query_recombinant', step=20, thread=1, win
dow=100, y_start=0.0)
VirusRecom starts calculating weighted information content from each lineage...


>>> Treat query_recombinant as a potential recombination lineage...
The calculation of reference_lineage_1's recombination contribution to query_recombinant has been completed!

The calculation of reference_lineage_2's recombination contribution to query_recombinant has been completed!

The calculation of reference_lineage_3's recombination contribution to query_recombinant has been completed!
```

(iii) At the end of the run, a concise report is printed reporting recombination events with p-values less than 0.05.

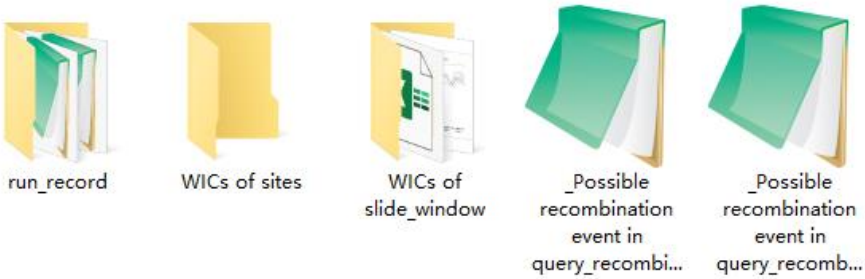
```
Possible major parent: reference_lineage_1(global mWIC: 1.8976186779157704)

Other possible parents:

Possible recombination region map at aligned genomes:

reference_lineage_2 [['7237 to 11539(mWIC: 1.9553354371515168)', 'p_value: 7.831109305531908e-06']]
```

(iv) The summary file of this result is at Possible recombination event in

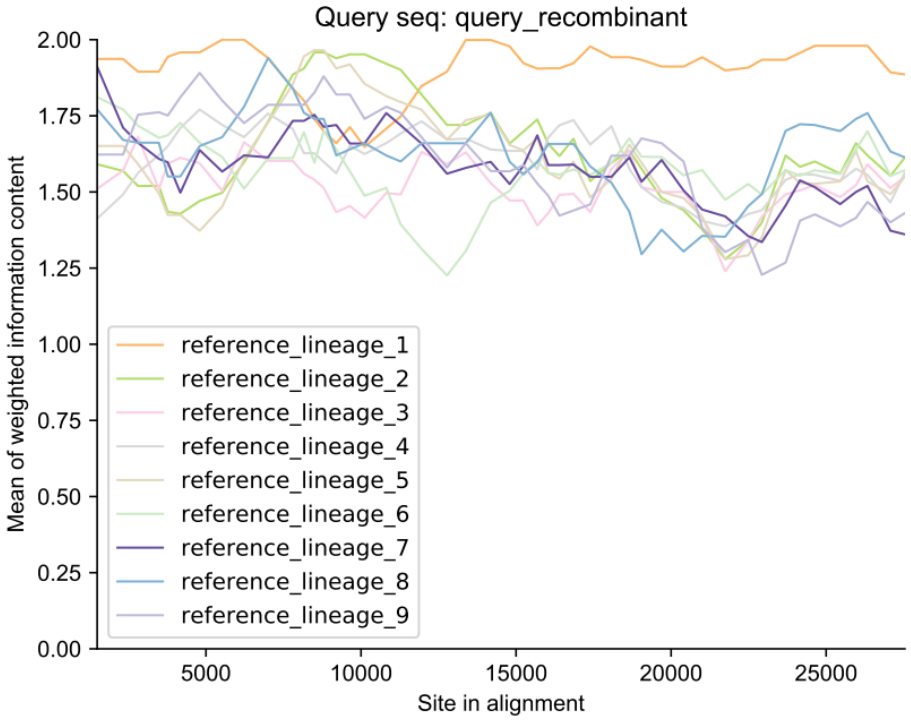
query_recombinant_conciseness.txt.



The intermediate files generated during the running process are located in the run_record directory.



(v) The matrix data generated by the sliding window operation is located in the directory of WICs of slide_window/_query_recombinant_WIC contribution from lineage in sliding window.xlsx. Users can use these raw data to draw graphs. In fact, virusrecom comes with a drawing function and provides drawn graphics (WICs of slide_window/_query_recombinant_WIC contribution from lineage in sliding window.pdf), and they might be as follows.



If the user thinks that the color of this picture is not good, the user can use the original matrix data provided by virusrecom to redraw.

6. Bug report

You can tell us any problems which you encounter in usage, and so that we can further improve VirusRecom. Submit your question in GitHub issuse (https://github.com/ZhijianZhou01/VirusRecom/issues) of VirusRecom or send email to zjzhou@hnu.edu.cn.

References

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol, 30*(4), 772-780. doi:10.1093/molbev/mst010