Manual of VirusRecom

Version 1.1

Detecting recombination of viral lineages using information theory

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

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

Version 1.1 || Mar 07, 2023

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 the recombination region is homologous or similar. However, aberrant homologous generally results in sequence deletion or duplication due to the recombination is not in the homologous region.

Identifying homologous recombination 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 quantifying recombination contribution using weighted information content (WIC). The For algorithm details, see (Zhou et al., 2023)

1.2. Functions

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

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

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

(III) Scan the potential recombination breakpoint.

2. Download and install

VirusRecom's code and all the updated versions are freely available at https://github.com/ZhijianZhou01/virusrecom. Due to cross-platform of Python, users can directly run the code or release package in Windows, MacOS or Linux systems.

For the release package, the executable file is located at the folder of *Main*, which is **virusrecom.exe** (windows system) or **virusrecom** (Linux or MacOS system). If you don'n own the permission to run virusrecom on Linux or MacOS system, you could change permissions by '*chmod -R 775 directory_path*' or '*chmod -R 777 directory_path*'.

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*****.**

Tip: virusrecom v1.1 optimizes the parameters of input-file, which is slightly different from virusrecom v1.0.

```
virusrecom [-h] [-a ALIGNMENT] [-ua UNALIGNMENT] [-at ALIGN_TOOL]
[-iwic INPUT_WIC] [-q QUERY] [-l LINEAGE] [-g GAP] [-m METHOD] [-w WINDOW]
[-s STEP] [-mr MAX_REGION] [-cp PERCENTAGE] [-cu CUMULATIVE]
[-b BREAKPOINT] [-bw BREAKWIN] [-t THREAD] [-y Y_START] [-le LEGEND]
[-owic ONLY_WIC] [-e ENGRAVE] [-en EXPORT_NAME] [-o OUTDIR]


☆ Example of use ☆
   (1) If the input sequence-data has been aligned:
   virusrecom -a alignment.fasta -q query_name -l lineage_name_list.txt -g n -m p -w 100 -s
20 -o outdir
```

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

```
virusrecom -ua unalignment.fasta -at mafft -q query_name -l lineage_name_list.txt -g n -m p -w 100 -s 20 -t 2 -o outdir
```

3.2. Parameters

Parameters	**Description**
-h, --help	Show this help message and exit.
-a ALIGNMENT	FilePath of an aligned sequence set (*.fasta format) containing all sequences used for analysis. **Tip**: the name of each sequence requires containing the lineage mark (a unique string).
-ua UNALIGNMENT	FilePath of a unaligned (non-aligned) sequence set (*.fasta format) containing all sequences used for analysis. **Tip:** the name of each sequence requires containing the lineage mark (a unique string).
-at ALIGN_TOOL	Program for multiple sequence alignments (MSA) if *-ua UNALIGNMENT* is used. Supporting mafft, muscle and clustalo, such as '*-at mafft*' or '*-at muscle*'.
-iwic INPUT_WIC	Using the already obtained WIC values of reference lineages directly by a *.csv input-file. Due to the big overhead of calculating WIC, thus, it can be reused.
-q QUERY	Name of query lineage (usually potential recombinant), such as '*-q xxxx*'. Besides, '*-q auto*' can scan all lineages as potential recombinant in turn.
-l LINEAGES	Path of a text-file containing the marks of reference lineages. **Tip: if '*-q auto*' is used**, the text-file needs to contain the marks of the query lineage and reference lineages.
-g GAP	Reserve sites containing gap in subsequent analyses? '*-g y*' means to reserve, and '*-g n*' means to delete.
-m METHOD	Scanning method of recombination analysis. '*-m p*' means use polymorphic sites only, '*-m a*' means all the monomorphic sites and

	polymorphic sites.
-w WINDOW	Number of nucleotides sites per sliding window. Note: if the '*-m p*' has been used, *-w* refers to the number of polymorphic sites per windows.
-s STEP	Step size of the sliding window. Note: if the '*-m p*' has been used, *-s* refers to the number of polymorphic sites per jump.
-mr MAX_REGION	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.
-cp PERCENTAGE	The cutoff threshold of proportion (*cp*, default value is 0.9) used for searching recombination regions when mWIC/EIC >= *cp*, the maximum value of *cp* is 1.
-cu CUMULATIVE	Simply using the max cumulative WIC of all sites to identify the major parent. The default value is 'n' and means off. If required, please specify '*-cm y*'.
-b BREAKPOINT	Whether to run the breakpoint scan of recombination. '*-b y*' means yes, '*-b n*' means no. Note: this option only takes effect when '*-m p*' has been specified.
-bw BREAKWIN	The window size (polymorphic sites, default value is 200) used for breakpoint scan, and step size is fixed at 1. **Note:** this option only takes effect when '*-m p -b y*' has been specified.
-t THREAD	Number of threads used for the multiple sequence alignments (MSA), and the default value is 1.
-y Y_START	Specify the starting value of the Y-axis in plot diagram, and the default value is 0.
-le LEGEND	The location of the legend, the default is adaptive. '*-le r*' indicates placed on the right.
-owic ONLY_WIC	Only calculate site WIC value in analysis. Off by default. If required, please specify '*-owic y*'.

-e ENGRAVE	Engraves file name to sequence names in batches. By specifying a directory containing one or multiple sequence files (*.fasta). It can be used to mark lineage when preparing data.
-en EXPORT_NAME	Export all sequence name of a *.fasta file.
-o OUTDIR	Output directory to store all results.

4. Example of usage

The sequences data for test of VirusRecom was stored at https://github.com/ZhijianZhou01/virusrecom/tree/main/example.

Note, the *recombination_test_data.zip* in directory *example* is against virusrecom v1.0, not virusrecom v1.1.

In this demonstration, the test data is from the the *recombination_test_data_v1.1.zip* provided in the directory *example*.

4.1. Aligned input-sequences

If the input sequence-data has been aligned, and it should be loaded via the *-a* parameter. Multiple sequence alignments (MSA) can be pre-completed by many programs, this is not introduced. Now, let's focus on the directory *aligned_input_sequences* in the file *recombination_test_data_v1.1.zip*.

(1) An aligned sequence-file named *alignment_lineages_data.fasta*, which including multiple sequences from the query lineage and other reference lineages.

(2) A text-file named *reference_lineages_name.txt*, which including the names (marks) of these reference lineages.

```
reference_lineage_1
reference_lineage_2
reference_lineage_3
reference_lineage_4
```

```
reference_lineage_5
reference_lineage_6
reference_lineage_7
reference_lineage_8
reference_lineage_9
```

Note, these marks of reference lineages should also appear in sequence names of the file *alignment_lineages_data.fasta*. **The mark of each reference lineage should be unique**, otherwise, there will be duplicate matches in subsequent analysis.

Before running VirusRecom, let's think about the search strategy for recombination events.

Firstly, we use only polymorphic sites considering that sequences from these lineages are highly similar, which means that the parameter *-m p* needs to be specified. Secondly, we do not consider gap-containing sites in this test and use the parameter *-g n*. Instead, if you consider these gap sites, you need to use the parameter *-g y*. Next, in the first run, let's try first with a window size of 100 and a step size of 20. Of note the value of "size" at this time represents the number of polymorphic sites because the *-m p* parameter has been specified. For the two parameters *-cp* and *-mr*, we use the default value of 0.9 and 1000 in this test. Finally, we specify a directory to save the results by parameter *-o*.

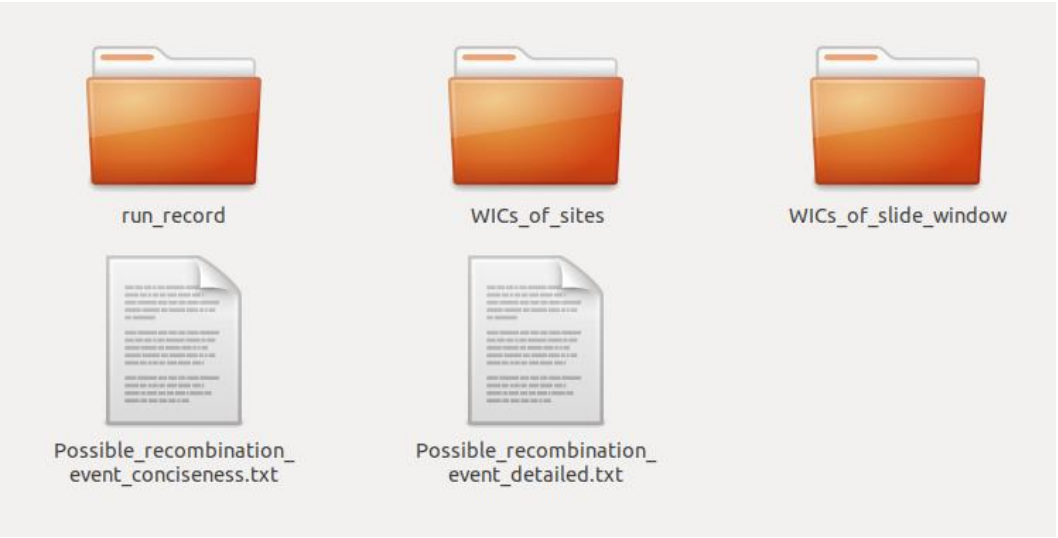
Then, switch the current directory to *aligned_input_sequences*, and run the following command (an example) to detect recombination events in query lineage:

```
virusrecom -a alignment_lineages_data.fasta -q query_recombinant -l
reference_lineages_name.txt -g n -m p -w 100 -s 20 -o outdir
```

Note: (1) if the current directory is not switched to *aligned_input_sequences*, the file and directory path in command need the absolute paths instead of relative paths.

(2) the string "query_recombinant" in command is the corresponding mark of query lineage in the file *alignment_lineages_data.fasta*.

After the run is complete, in the directory *outdir*, there are three subdirectories and two aggregated reports:



(1) In the directory *run_record*, if *-g n* is specified, and the file *Record_of_deleted_gap_sites_*.txt* containing all the gap sites will be created. Besides, If *-m p* is specified, and the file *Record_of_same_sites_in_aligned_sequence*.txt* containing all the same sites will be created.

(2) In the directory *WICs_of_sites*, the file **_site_WIC_from_lineages.pdf*, **_site_WIC_from_lineages.xlsx* and the file **_site_WIC.csv* are used to record the WIC value of each site.

(3) In the directory *WICs_of_slide_window*, the file **_mWIC_from_lineages.xlsx* and the file **_mWIC_from_lineages.pdf* are used to record the mean WIC of each sliding window.

The user can fine-tune the window size and step size according to the density of points in the generated graph. In general, very dense points means that the noise is too high and the window size can be increased appropriately in next scan.

In addition to the three sub-directories above, VirusRecom provides two summary files. The file *Possible_recombination_event_conciseness.txt* only retains results of recombination events with p-values less than 0.05.

Possible major parent: reference_lineage_1(global mWIC: 1.8976186779157704)

Other possible parents and significant recombination regions (p<0.05):
reference_lineage_2 7237 to 11539(mWIC: 1.9553354371515168), p_value: 7.831109305531908e-06

In this output report, the major parent of query lineage was *reference_lineage_1* and the minor parent was *reference_lineage_2*, and the recombination region was site 7237 to 11539 and the p-value was 7.83e-06. The identified recombination event was relatively close to the actual (from site 7333 to 11473 in the genome), and the error of

the recombination boundary is also acceptable.

In fact, *Possible_recombination_event_conciseness.txt* is interpretations of the recombination information contained in **_mWIC_from_lineages.pdf*. Although VirusRecom shows a good balance between precision and recall in simulated data, false positive or false negatives sometimes occur. Therefore, for the identification results from VirusRecom, users can make own judgment.

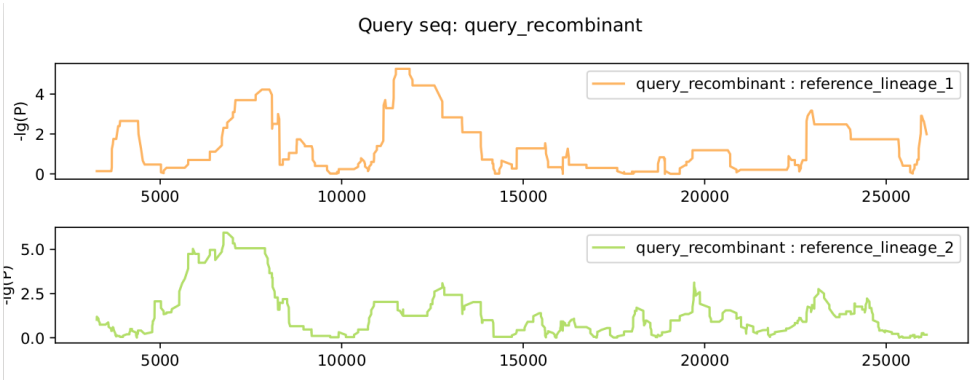
Besides, the output file *Possible_recombination_event_detailed.txt* shows those results with p-values greater than 0.05. Tip: recombination events with p-values over 0.001 are less reliable.

If *-b y* is specified, then VirusRecom will perform the search of recombination breakpoint and plot. For example:

```
virusrecom -a alignment_lineages_data.fasta -q query_recombinant -l
reference_lineages_name.txt -g n -m p -w 100 -s 20 -b y -bw 200 -o outdir
```

Tip: (1) *-b y* only takes effect when '*-m p*' has been specified. (2) the step size of breakpoint search is fixed to 1.

The negative logarithm of *p*-value in each site is in the file **_-lg(p-value)_for_potential_breakpoint.pdf* and the file **_-lg(p-value)_for_potential_breakpoint.xlsx*.



The highest peak (the highest −lgP value) indicated the possible recombination breakpoint.

4.2. Unaligned input-sequences

VirusRecom can also handle unaligned input-sequences. In this case, multiple sequence alignment is performed by calling external program. **In virusrecom v1.1**, mafft, muscle, and clustal-omega is supported (Edgar, 2004; Katoh & Standley, 2013; Thompson, Higgins, & Gibson, 1994). It is worth mentioning that VirusRecom call them from the system path, so they need to be installed on the machine beforehand.

For the example data in directory *unaligned_input_sequences*, run the following command:

```
virusrecom -ua unalignment_lineages_data.fas -at mafft -q query_recombinant -l reference_lineages_name.txt -g n -m p -w 100 -s 20 -o outdir
```

Note: (1) *"-at mafft"* means to call mafft in the system path, and the alignment strategy is *auto*. Besides, using *"-at muscle"* to call muscle and using *"-at clustalo"* to call clustal-omega.

(2) the string "query_recombinant" in command is the corresponding mark of query lineage in the file *unalignment_lineages_data.fas*.

The interpretation of the output result is consistent with section 4.1.

5. Reuse site WIC value

Throughout the analysis, getting site WIC is a very time-consuming calculation. In practice, we may need to adjust the parameters to hunt for recombination events, such window size (*-w*), step size (*-s*) and cutoff threshold of proportion (*-cp*) and so on.

In virusrecom v1.1, users can use the already obtained WIC values of reference lineages directly by a *.csv input-file. The example is as follows:

```
virusrecom -iwic query_recombinant_site_WIC.csv -q query_recombinant -g n -w 100 -s 20 -o outdir
```

Note, *"-g n"* is used here because *query_recombinant_site_WIC.csv* was previously calculated under *-g n*. The example data *query_recombinant_site_WIC.csv* is in *wic_file* directory.

6. Non-lineage data

In VirusRecom, the reference lineage is allowed to contain only one single sequence. Under this condition, mWIC value of the fragment is essentially a multiple of shared identity. If *-g n* is used in the calculation, the mWIC is twice as large as shared identity. If *-g y* is used in the calculation, the mWIC is $\log_2 5$ as large as shared identity.

Of noted, for recombination analysis without lineage data, the additional feature is only recommended for non-highly similar sequences and the user can use it to draw an identity point map.

The test data is in directory *non_lineage_data* of *recombination_test_data_v1.1.zip*.

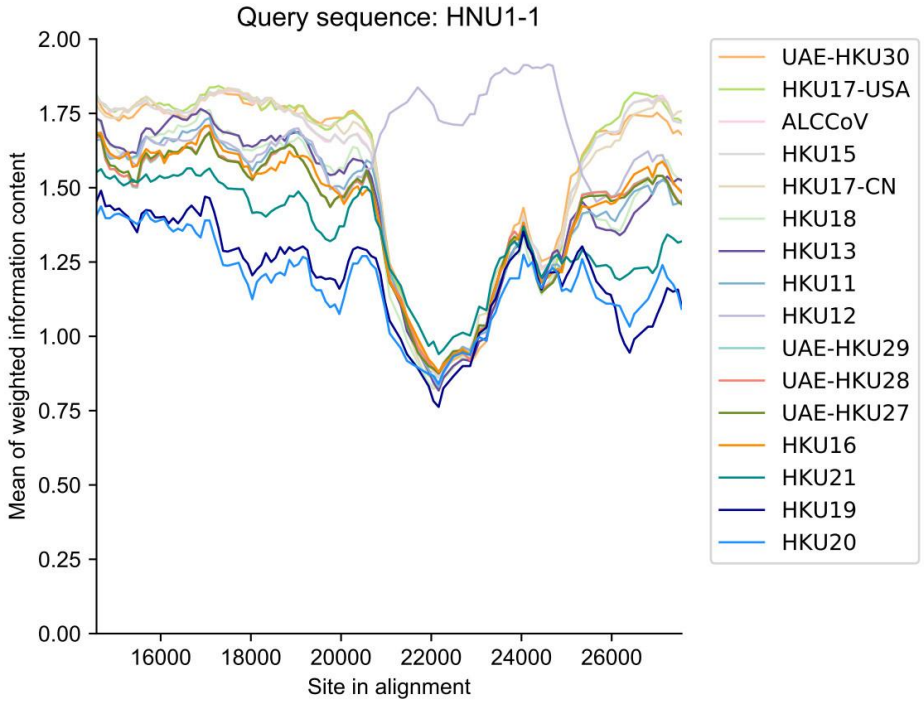
The Delta-CoV HNU1-1 is a known recombinant from SpCoV HKU17-USA and ThCoV HKU12, and the break points were identified at genome positions nt 21017 and 25056, which is jointly identified and confirmed by RDP3 and Simplot (Wang et al., 2022).

Considering that they are not highly similar sequences, we use all sites (*-m a*) in the

alignment. Then, we use a larger window value, and run following command:

```
virusrecom -a alns.fasta -q HNU1-1 -l alns_seq_taxon.txt -g n -m a -w 800 -s 100 -cp 0.7
-mr 6000 -le r -o output
```

The mWIC from reference lineages is as follows:



Note, because each "lineage" contains only one sequence and *-g n* is used in the example, the mWIC in the picture is actually twice the size of "sequence identity".

The possible recombination event identified by VirusRecom is as follows:

```
Possible major parent: HKU17-USA(global mWIC: 1.5914816042426252)

Other possible parents and significant recombination regions (p<0.05):
HKU12     20720 to 25297(mWIC: 1.8039433490697028), p_value: 2.783880536189705e-204
```

The possible major parent of HNU1-1 is HKU17-USA and minor parent is HKU12, and the recombination region is about 20720-25297 nt in the alignment.

7. Common questions

7.1. Default values of parameter

For the value of a parameter, if not specified, the software uses the default value. However, the default value is not suitable for all data. In addition to window size (-*w*) and step size (-*s*) of sliding window, values of -*cp* and -*mr* also require users to adjust based on the data.

When VirusRecom runs, the value of each parameter is printed printed on the screen and you can check them.

7.2. How to mark lineage in sequence name?

Typically, this is part of the data preparation. In virusrecom v1.1, users can easily get it done via "-*e*" parameter. The "-*e*" parameter can engrave file-name to sequence names in batches. The example is as follows:

```
virusrecom -e input_directory -o outdir
```

Tip: The directory "input_directory" can contain multiple fasta files, and each fasta file can contain multiple sequences. After the running, finally, each sequence name will contain its file-name.

Therefore, if the file-name of fasta file is a lineage name, the lineage name can be written into the sequence name in batches.

7.3. How to change the color scheme in an image?

If you own programming skills, you can directly modify the order of the colors in the *plt_corlor_list.py* file. If not, you can use output matrix provided by VirusRecom, and

they are usually suffixed with .xlsx.

8. Bug report or requests

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

References

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

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

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res, 22*(22), 4673-4680. doi:10.1093/nar/22.22.4673

Wang, Q., Zhou, Z. J., You, Z., Wu, D. Y., Liu, S. J., Zhang, W. L., . . . Ge, X. Y. (2022). Epidemiology and evolution of novel deltacoronaviruses in birds in central China. *Transbound Emerg Dis, 69*(2), 632-644. doi:10.1111/tbed.14029

Zhou, Z. J., Yang, C. H., Ye, S. B., Yu, X. W., Qiu, Y., & Ge, X. Y. (2023). VirusRecom: an information-theory-based method for recombination detection of viral lineages and its application on SARS-CoV-2. *Brief Bioinform, 24*(1). doi:10.1093/bib/bbac513