The Journal of Open Source Software

Recan: Python tool for analysis of recombination events in viral genomes

Yuriy Babin[1]

1 National Medical Research Center for Tuberculosis and Infectious Diseases, Moscow, Russia

DOI: 10.21105/joss.02014

Software

- Review
- Repository
- Archive

Editor: William Rowe

Reviewers:

- @lamhm
- @betteridiot

Submitted: 20 December 2019
Published: 12 May 2020

Summary

Recombination drives virus evolution in response to selective forces in a host environment and adaptation to new abiotic factors (Pérez-Losada, Arenas, Galán, Palero, & González-Candelas, 2015). Gaining insights into recombination events is important for a better understanding of viral biology. Analysis of recombination events can be performed through construction and exploration of similarity plots based on genetic distances between nucleotide sequences. Python package named "recan" (recombination analyzer) provides the means to construct genetic distance plots and explore them interactively. The package has been designed to operate with the Jupyter notebook. Compared to the previously designed desktop software (Etherington, Dicks, & Roberts, 2005; Lole et al., 1999; Martin, Murrell, Golden, Khoosal, & Muhire, 2015) recan has the ability to insert or delete sequences from the output without reconstructing the plots and recalculating the distance values. Finally, recan enables simultaneous analysis of several datasets in a single session. Recan is based on Biopython, Pandas, and Plotly libraries. The package requires a sequence alignment in fasta format as an input. The user can adjust the sliding window size, the window shift, method of distance calculation, sequence of interest (a sequence where breakpoints occur), and the length of alignment region which will be included into the distance calculation. The two methods of genetic distance calculations implemented in recan are the pairwise and Kimura 2-parameter models. The distance data can be saved in csv or excel file, or directly used to reconstruct the plot in the Jupyter notebook using the plotting library to obtain a final report.

Testing and verification

To test the package, we used four previously reported recombinant viral genomes representing different genuses: human immunodeficiency virus (HIV) (Liitsola et al., 2000), hepatitis C virus (HCV) (Smith et al., 2014), norovirus (Jiang, Espul, Zhong, Cuello, & Matson, 1999), and lumpy skin disease virus (LSDV) (Alexander Sprygin, 2018). Each dataset included a recombinant virus sequence, its putative parental sequences and a set of sequences of the same virus closely related to the recombinant virus. HIV, HCV and Norovirus sequences were aligned using ClustalW (Larkin et al., 2007), and LSDV genomes were aligned using MAFFT (Katoh, 2002) as part of Ugene software (Okonechnikov et al., 2012). The HIV alignment contained twenty five 3135 bp sequences; the HCV alignment contained twenty three 9431 bp sequences; the norovirus alignment included nineteen 3366 bp sequences, and the LSDV alignment had a total of 150511 bp sequences. The resulting `simgen` method execution time with the default window size and shift parameters was the following: 437 ms $\pm$ 7.74 ms for HIV, 579 ms $\pm$ 58.7 ms for Norovirus, 648 ms $\pm$ 44.2 ms for HCV, and 3.55 s $\pm$ 239 ms for LSDV dataset. Time execution test was performed using a desktop PC with 4 CPU cores and 4 Gb RAM. LSDV has one of the largest genomes of all viruses (about 150 000 bp).

Ultimately, recan can potentially be used to identify and analyze recombination events in a large subset of sequences regardless of the length of the viral genome. The distance plots with recombination events detected by recan are shown in Figures 1-4.

Availability and implementation

Recan is supported on Linux and Windows. The package can be installed by `pip` Python package manager using `pip install recan` command. The source code, guide and datasets are available on the GitHub repository (https://github.com/babinyurii/recan).

Figure 1. HIV recombinant strain AF193276 between sequences AF193275 and AF193278.

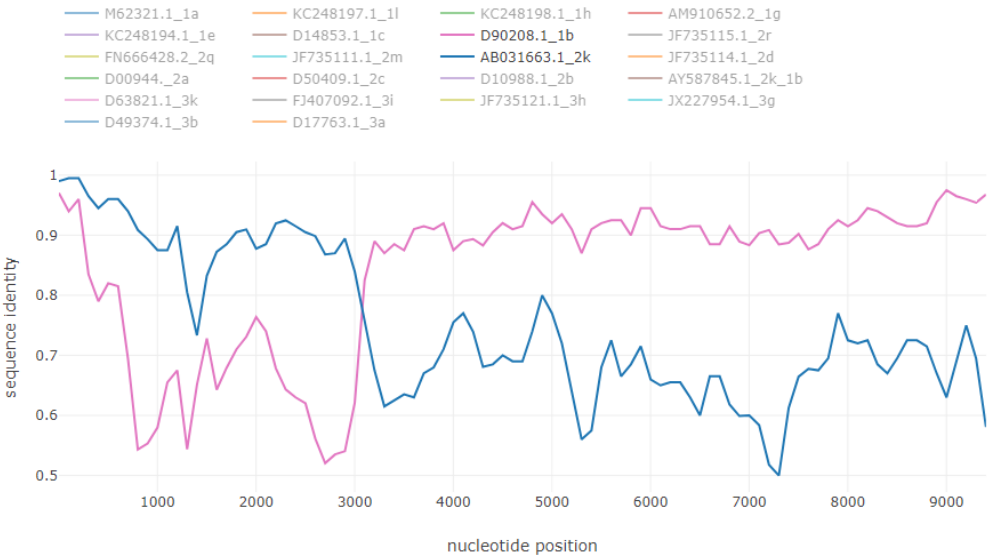

Figure 2. HCV intergenotype recombinant 2k/1b.

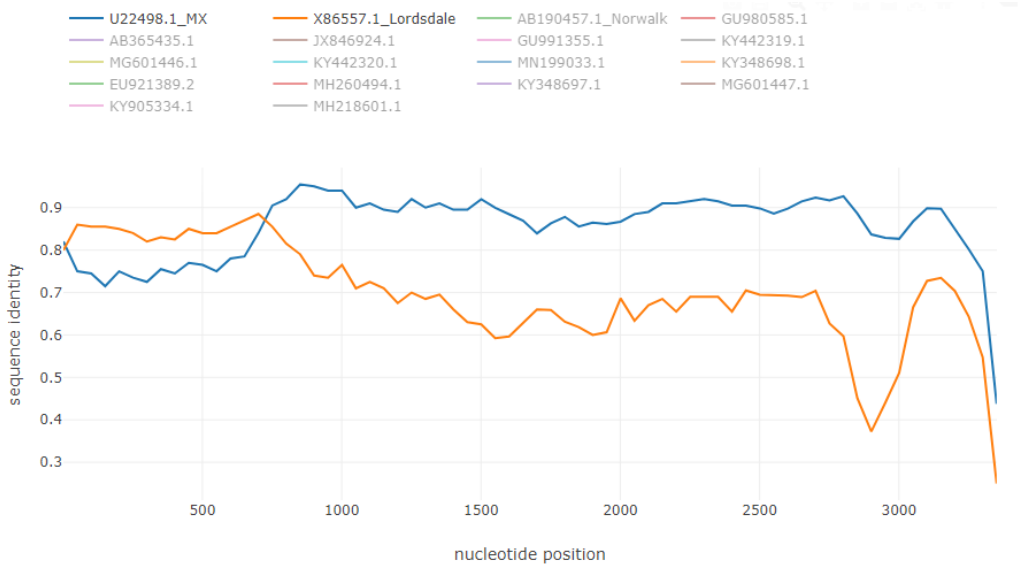

Figure 3. Norovirus recombinant AF190817 between parental sequences U22498 and X86557.

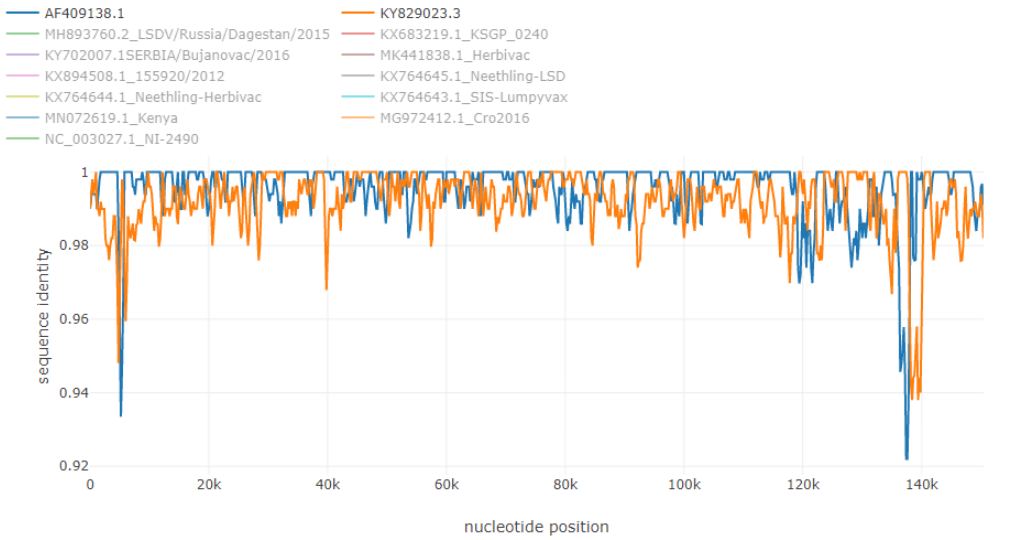

Figure 4. LSDV recombinant vaccine-like strain LSDV RUSSIA/Saratov/2017 between sequences AF193275 and KY829023.

Acknowledgement

The author thanks Alexander Sprygin for editing the manuscript and providing LSDV data.

References

Alexander Sprygin, Y. P., Yurii Babin. (2018). Analysis and insights into recombination signals in lumpy skin disease virus recovered in the field. *PLoS ONE*, *13*(12), 1–19. doi:10.1371/journal.pone.0207480

Etherington, G. J., Dicks, J., & Roberts, I. N. (2005). Recombination Analysis Tool (RAT): A program for the high-throughput detection of recombination. *Bioinformatics*, *21*(3),

278–281. doi:10.1093/bioinformatics/bth500

Jiang, X., Espul, C., Zhong, W. M., Cuello, H., & Matson, D. O. (1999). Characterization of a novel human calicivirus that may be a naturally occurring recombinant. *Archives of Virology*, *144*(12), 2377–2387. doi:10.1007/s007050050651

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. doi:10.1093/nar/gkf436

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., Valentin, F., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. doi:10.1093/bioinformatics/btm404

Liitsola, K., Holm, K., Bobkov, A., Pokrovsky, V., Smolskaya, T., Leinikki, P., Osmanov, S., et al. (2000). An AB recombinant and its parental HIV type 1 strains in the area of the former Soviet Union: Low requirements for sequence identity in recombination. *AIDS Research and Human Retroviruses*, *16*(11), 1047–1053. doi:10.1089/08892220050075309

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of virology*, *73*(1), 152–60.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1), 1–5. doi:10.1093/ve/vev003

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., German Grehov, O. G., et al. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166–1167. doi:10.1093/bioinformatics/bts091

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, *30*(December), 296–307. doi:10.1016/j.meegid.2014.12.022

Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T., & Simmonds, P. (2014). Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: Updated criteria and genotype assignment web resource. *Hepatology*, *59*(1), 318–327. doi:10.1002/hep.26744