

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

Yuriy Babin¹

¹ National Medical Research Center for Tuberculosis and Infectious Diseases

DOI: [10.21105/joss.02014](https://doi.org/10.21105/joss.02014)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [William Rowe](#) ↗

Reviewers:

- [@lamhm](#)
- [@betteridiot](#)

Submitted: 20 December 2019

Published: 14 January 2020

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

Recombination is widespread in viruses and is considered to be one of the mechanisms that drives evolutionary changes and helps viruses to overcome selective pressure and adapt to new environments (Pérez-Losada, Arenas, Galán, Palero, & González-Candelas, 2015). The similarity plot based on genetic distance between nucleotide sequences is an intuitive and simple way to explore and discover recombination events in viral genomes. Python package named recan (recombination analyzer) allows to construct genetic distance plots and explore them interactively. Recan is intended to be used via Jupyter notebook which is a popular instrument for data analysis. The aim of the package is to give the researchers who are familiar with Python a tool which allows to perform recombination analysis easily and quickly. The advantages of recan above the distance methods implemented previously in desktop software (Etherington, Dicks, & Roberts, 2005; Lole et al., 1999; Martin, Murrell, Golden, Khoosal, & Muhire, 2015) are its interactivity, speed, the ability to easily include sequences into or exclude them from the output without reconstructing the plot and recalculating the distance data, and, at last, the ability to analyze simultaneously several datasets during a single session. Recan is based on Biopython, Pandas, and Plotly libraries. The package requires sequence alignment in fasta format as an input. A user can control the sliding window size, the window shift, a method of distance calculation, a sequence of interest (which is a sequence potentially containing recombination region), and the alignment region which will be included into distance calculation. Two methods of genetic distance calculation implemented so far are pairwise and Kimura 2-parameter distance. After a recombination event is found, the distance data can be saved in csv or excel file, or directly used to reconstruct the plot in Jupyter notebook with ones favorite plotting library to get a final report.

Testing and verification

To test the package four previously reported recombinant viral genomes were used: 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 sequence, its parental sequences and a number of other sequences of the same virus closely related to the recombinant. HIV, HCV and Norovirus sequences were aligned with Clustal W (Larkin et al., 2007) and LSDV genomes were aligned with MAFFT (Katoh, 2002) by using Ugene software (Okonechnikov et al., 2012). Finally, HIV alignment included 25 sequences of the length 3135 bp, HCV alignment had 23 sequences of the length 9431, Norovirus alignment included 19 genomes of the length 3366, and LSDV dataset had 14 sequences of the length 150511 bp. Recan `simgen` method execution time with the default `window` 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 $3.55 \text{ s} \pm 239 \text{ ms}$ for LSDV dataset. Time execution test was performed using the desktop PC which has 4-core CPU and 4 Gb RAM. LSDV genome is about 150 000 bp long and is one of the largest among viruses. Typically researches study recombination events using the viruses which genomes are way smaller. But even in the case of analyzing the dataset which includes up to 14 large viral genomes recan execution time is quite reasonable. Distance plots with recombination events detected by recan in the test datasets are shown in Figures 1-4.

Availability and implementation

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

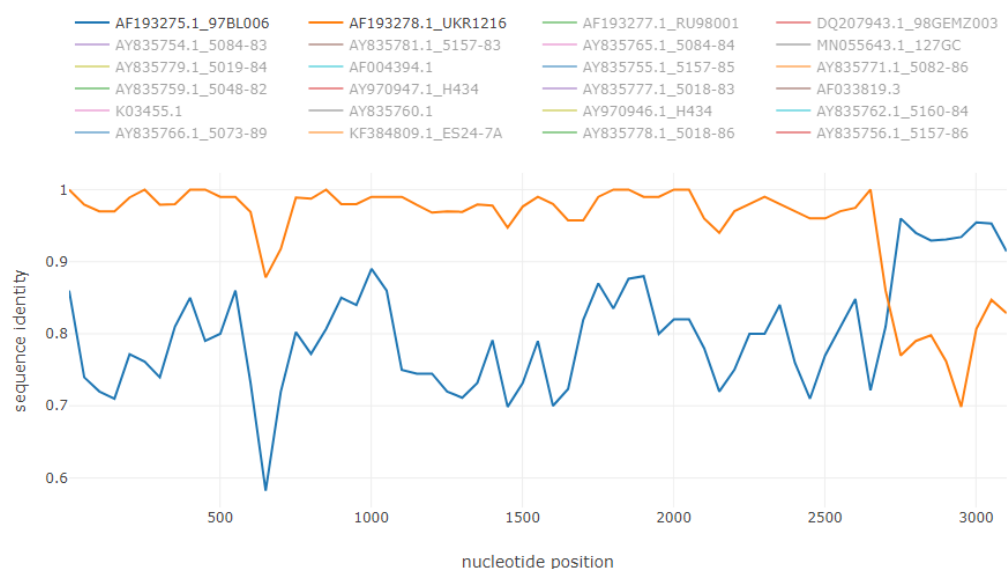


Figure 1. HIV recombinant strain AF193276 between sequences AF193275 and AF193278.



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

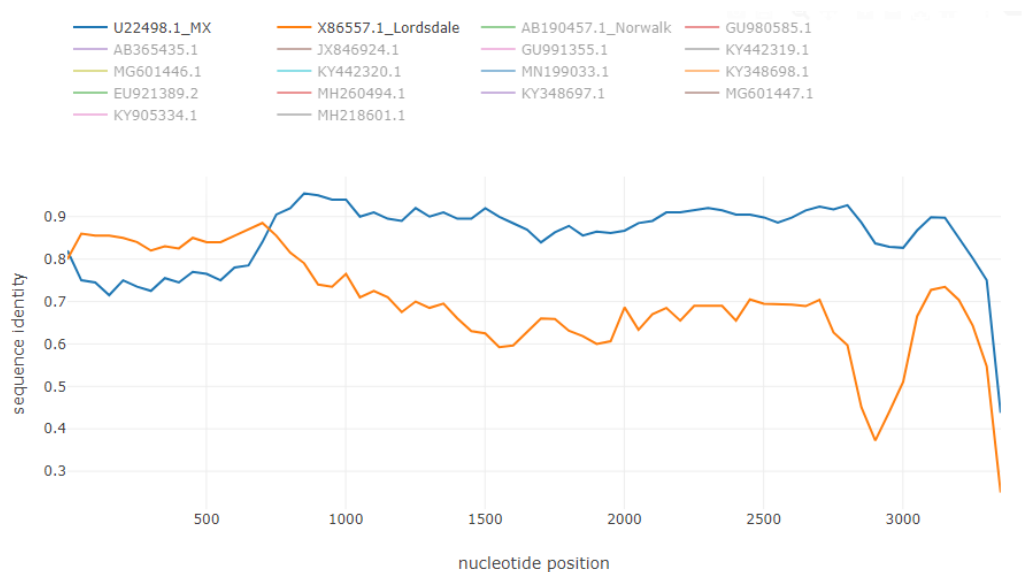


Figure 3. Norovirus recombinant AF190817 between parental sequences U22498 and X86557.

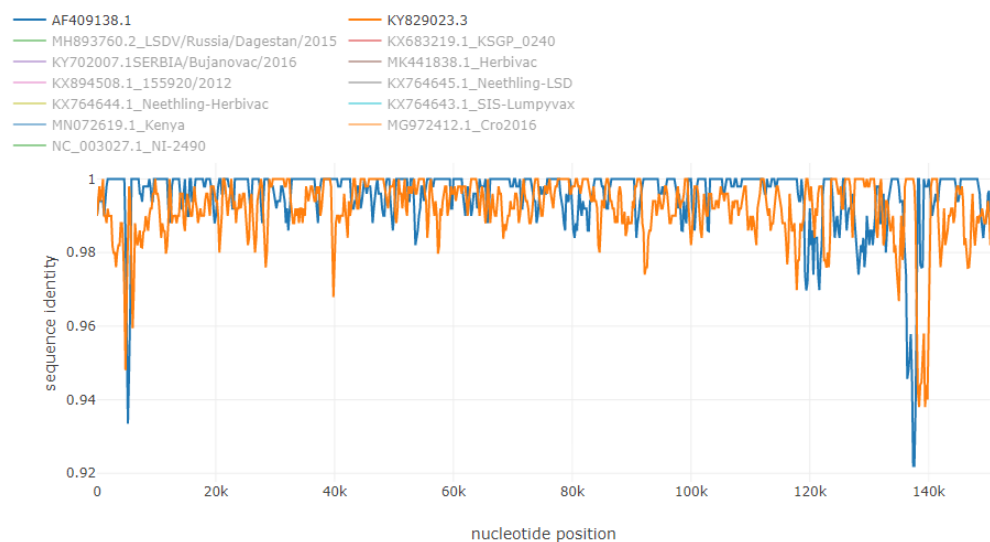


Figure 4. LSDV recombinant vaccine-like strain LSDV RUSSIA/Saratov/2017 between sequences AF193275 and KY829023.

References

- Alexander Sprygin, Y. P., Yuri 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1002/hep.26744)