Using machine learning to identify ancient RNA viruses in the human genome
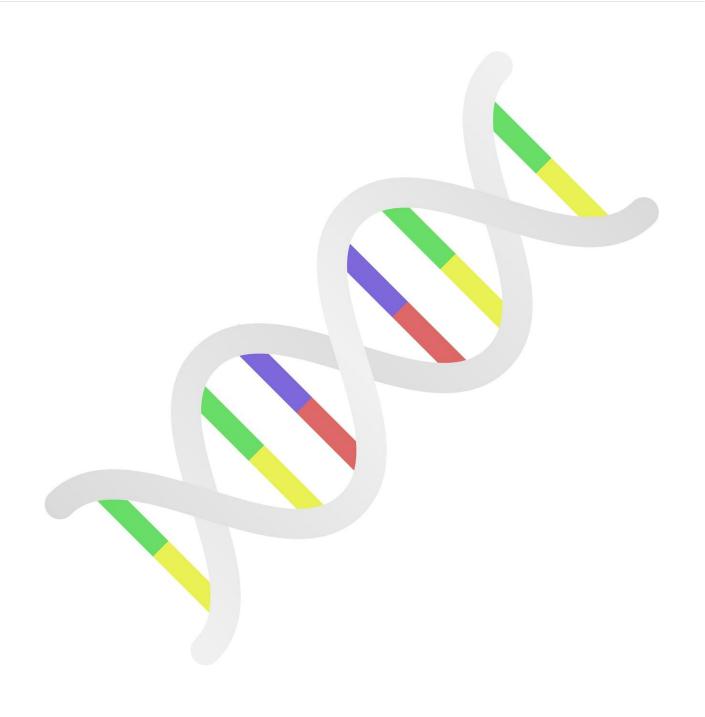
January 26 2021, by Bob Yirka

A team of researchers affiliated with multiple institutions in Japan has used machine learning algorithms to help them identify ancient RNA virus remnants in the human genome. In their paper published in *Proceedings of the National Academy of Sciences*, the group describes how they taught their AI system to recognize RNA virus remnants and then used it to scan the human genome.

Prior research has shown that when a person (or other animal) is infected with a virus, that virus can sometimes change the host's DNA by adding some of its own RNA. Other prior research has shown that ancient viruses that infected populations many years ago have sometimes left remnants of their RNA in the human genome. Finding such remnants has proved challenging, however, due to the huge numbers of comparisons required for each suspected virus. In this new effort, the researchers used a machine- learning algorithm to help with the search.

To train the algorithm, the researchers used RNA from known non-retroviral endogenous RNA virus elements. The thinking was that by training the algorithm with modern viral RNA, the system could get a feel for what viral RNA looks like in general. Researchers believe that such commonalities likely existed in ancient RNA, as well. After training, the researchers fine tuned their system to prevent as many false positives as possible. They then set it to work on the human genome and identified approximately 100 possibilities. After studying the possibilities, the researchers found that many of them were already known and many also fell below a threshold they had set to serve as a pass/fail option. That left them with just one possible unknown virus remnant.

The researchers then looked to see if the same remnant appeared in the genomes of other species such as marmosets and chimpanzees, and found that was, indeed, the case. That finding suggested the insertion had occurred at least 43 million years ago—before the species had diverged.

The researchers suggest their approach could be expanded to include other types of viruses, noting that the approach would likely reveal other remnants. They further note that learning more about ancient viruses and how they impacted the human genome could provide insight into modern virus behaviors that have not yet been discovered.

More information: Shohei Kojima et al. Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome, *Proceedings of the National Academy of Sciences* (2021). DOI: 10.1073/pnas.2010758118