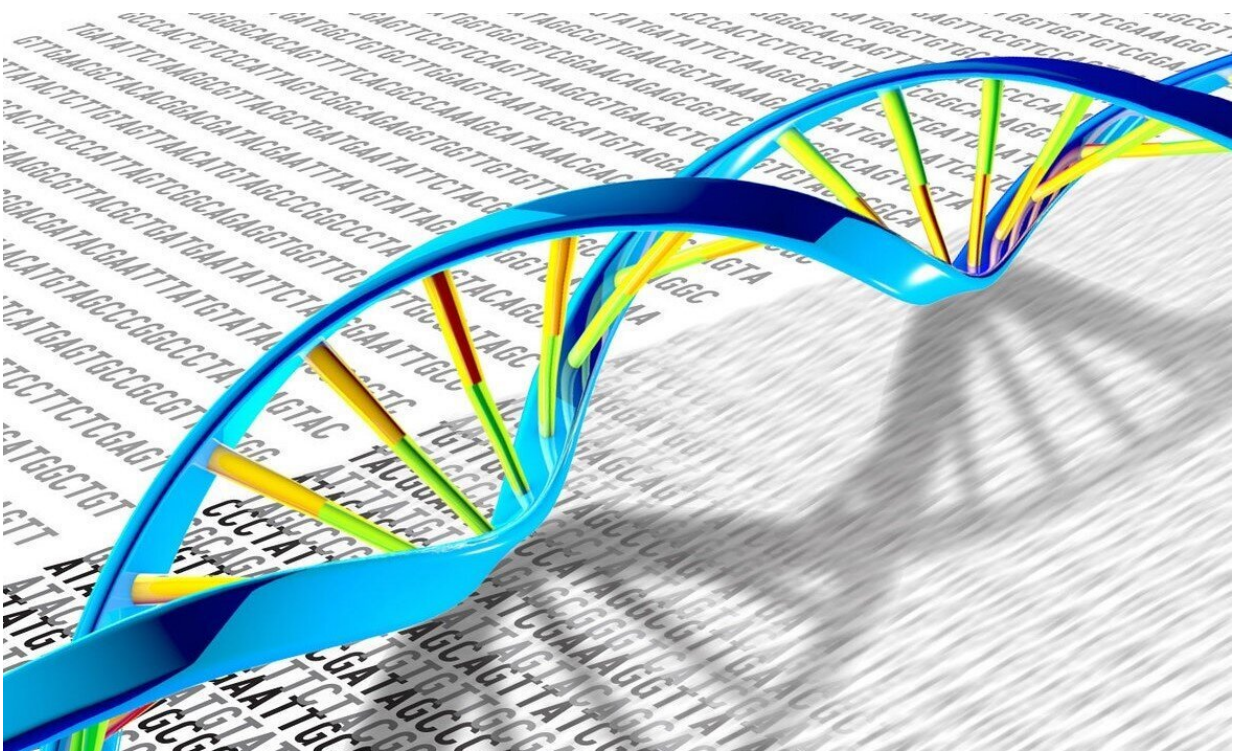# Machine learning identifies common DNA structures

July 9 2019

DNA, which has a double-helix structure, can have many genetic mutations and variations. Credit: NIH

Researchers from HSE University have used machine learning to discover that the two most widespread DNA structures—stem loops and quadruplexes—cause genome mutations that lead to cancer. The results of the study were published in *BMC Cancer*.

In the early 2000s, researchers invented a new method to obtain the nucleotide sequence of DNA and RNA, called Next-Generation Sequencing (NGS). This technology allows simultaneous reading of several million genome regions, which was impossible with earlier sequencing methods. Now, the human genome can be recorded in a 3.2 Gb text file.

"Cancer is a genome disease," explains Maria Poptsova, head of the HSE Laboratory of Bioinformatics and one of the study's authors. "When we sequence the genome in a tumour tissue, we see a spectrum of different mutations. There may be point or large-scale mutations. For example, in point mutations, one nucleotide disappears and is replaced by another. We looked at large-scale mutations where parts of the genome (from tens to millions of nucleotides) were deleted, reversed, copied, and inserted in a different place. As a result of these rearrangements, genome breakpoints appear.

Using machine learning, HSE University researchers investigated the influence of two types of DNA secondary structures—stem loops and quadruplexes—on genome breakpoints. The authors analysed a half-million breakpoints in over 2,000 genomes of 10 types of cancer. The researchers looked for genomic hotspots, considering breakpoint hotspots to be the regions with frequent and recurrent rearrangements—in other words, risk zones. It appeared that the stem loop-based model best explains blood, brain, liver and prostate cancer breakpoint hotspot profiles, while a quadruplex-based model has higher performance for bone, breast, ovary, pancreatic and skin cancer.

The appearance of breakpoints cannot be explained exclusively by the impact of DNA secondary structures, but their contribution is at least 20-30 percent. The analysis demonstrates that the impact of stem loops and quadruplexes on breakpoint evolution depends on the type of tissue, which is determined by epigenetic factors.

"These are the kind of markers that distinguish different kinds of tissues over the genome," said Maria Poptsova. "We are actively studying the correlation between secondary DNA structures and epigenetic marks. British researchers have already looked at the impact of DNA secondary structures and epigenetic marks on point mutations. We focused on breakpoint hotspots and are the first to determine the contribution of the two most widespread genome structures—stem loops and quadruplexes."

According to the study's authors, in the future, quadruplexes may be therapeutic targets. If drug therapy makes them more stable, the telomerase enzyme won't be able to work in cancer cells, and they will become vulnerable.

More information: Kseniia Cheloshkina et al, Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation, *BMC Cancer* (2019). DOI: 10.1186/s12885-019-5653-x

Provided by National Research University Higher School of Economics

Citation: Machine learning identifies common DNA structures (2019, July 9) retrieved 27 April 2024 from https://medicalxpress.com/news/2019-07-machine-common-dna.html