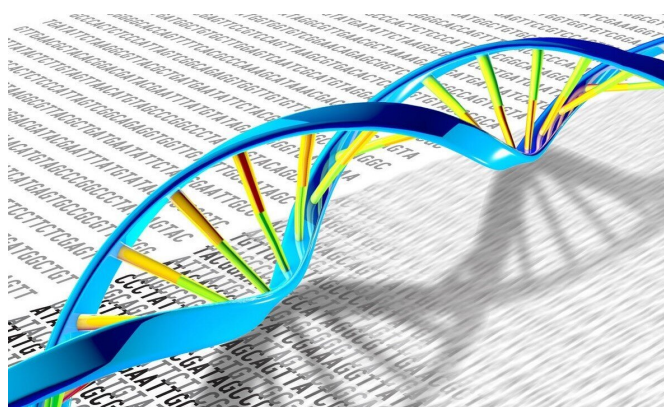# Researcher uses machine learning to demonstrate that DNA impacts cancer risk

23 December 2020

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

Lifestyle—or put another way, 'bad habits'—is one of the textbook explanations for why some people are at higher risk for cancer. We often hear that smoking increases our risk of developing lung cancer or that a high-fat diet increases our risk of developing bowel cancer, but not all smokers get lung cancer and not all people who eat cheeseburgers get bowel cancer. 'Other factors' must be at play.

Now, new research from University of Calgary scientist Dr. Edwin Wang, Ph.D., is shedding light on those 'other factors'. Wang has discovered seven DNA fingerprints or patterns that define cancer risk. The research is published in *Science Advances.*

"This discovery rewrites the textbook explanation that cancer occurs because of human behavior combined with some bad luck to include one's genetic make-up," says Wang. "We believe that a baby is born with a germline genomic pattern and it will not change, and that pattern is associated with a lower or higher cancer risk."

The research offers new insight into multi-generational disease risk as the germline represents the cells that determine our children and the DNA that is passed from parent to children. It is the first time scientists have described these highly-specialized biological patterns applicable to cancer risk.

Wang, a cancer systems biologist and big data scientist, holds the Alberta Innovates Translational Chair in Cancer Genomics. He hypothesizes that everyone fits into these risk categories making them more-or-less predisposed to cancer, much like a sliding scale. A member of the Alberta Children's Hospital Research Institute (ACHRI) and Arnie Charbonneau Cancer Institute at the Cumming School of Medicine, Wang found that the DNA fingerprints could be classified into subgroups with distinct survival rates. One of the seven germlines offers protection from developing cancer, and the other six germlines present a greater risk for cancer.

"It is interesting that one of these germlines is protective against developing cancer and it appeared frequently in our analysis of genomes," says Wang, a professor in the CSM's Department of Biochemistry and Molecular Biology. "We know there are individuals who can smoke and have an unhealthy lifestyle but never get cancer, and this discovery may explain that phenomena."

For this research, Wang conducted a massive systematic analysis of more than 26,000 germline genomes of individuals, about 10,000 people who had cancer, and the rest without. His team analyzed computer files from cancer patients at the National Cancer Institute—data collected by the National Institute of Health for the Cancer Genome Atlas, part of the National Institutes of Health in the U.S. The samples include 22 distinct cancers, including lung, pancreatic, bladder, breast, brain, stomach, thyroid, and bone and a dozen more. The control group of people without cancer included

genomic-sequenced groups from Sweden, England and Canada.

The massive quantities of data could only be processed with machine learning. Wang's lab is equipped to deal with data through ultra high-speed networks at UCalgary. This research requires a colossal amount of computer storage: 10 million terabytes. To help understand this volume, one terabyte can store 250 movies.

"Even at high-speed, with two streams running 24/7, it took our lab three straight months just to download the biological information containing billions and billions of nucleotides in each individual genome," says Wang.

Wang notes that between five to 10 percent of cancers are caused by specific gene mutations. Think of breast cancer and the inherited gene BRCA1 and BRCA2, a gene mutation made widely known by actor Angelina Jolie. Wang has always suspected these inherited cancers only represent a handful of associations and undertook a deeper investigation with advanced genomic capabilities to yield more associations.

"We wanted to investigate whether a genomic pattern or a substantial, repeatedly occurring sequential profile in genomes could serve as a promising measurement for genetic predisposition to cancer," says Wang.

"We found that one DNA-fingerprint was enriched tens to hundreds of times in germline genomes of cancer patients, suggesting that it is a universal inheritable trait encoding cancer risk." The research also uncovered that another DNA-fingerprint was highly enriched in cancer patients who were also tobacco smokers, indicating that smokers bearing such a DNA-fingerprint have a higher risk of cancer.

Genomic medicine makes diagnosis of disease more efficient, cost-effective, and can help people make health decisions throughout their life. Wang's research lays the groundwork for tools that could help cancer specialists and family physicians guide patients. "I hope that further studies are carried out to expand upon this work, so that it may eventually

be put into practice allowing clinicians to inform patients of their cancer risk and how to take precautions to ensure a healthy life."

Provided by University of Calgary