COLUMBIA | Zuckerman Institute

MORTIMER B. ZUCKERMAN MIND BRAIN BEHAVIOR INSTITUTE

Columbia Scientists Build Better Way to Decode the Genome

~ New computer algorithm deciphers DNA's most well-kept secrets, may help find the links between genes and disease ~

Date: April 6th, 2018
Contact: Anne Holden, anne.holden@columbia.edu, 212.853.0171

NEW YORK — The genome is the body's instruction manual. It contains the raw information — in the form of DNA — that determines everything from whether an animal walks on four legs or two, to one's potential risk for disease. But this manual is written in the language of biology, so making sense of all that it encodes has proven challenging. Now, Columbia University researchers have developed a computational tool that shines a light on the genome's most hard-to-translate segments. With this tool in hand, scientists can get closer to understanding how DNA guides everything from growth and development to aging and disease.

The researchers recently published their findings in the *Proceedings of the National Academy of Sciences*.

"The genomes of even simple organisms such as the fruit fly contain 120 million letters worth of DNA, much of which has yet to be decoded because the cues its provides have been too subtle for existing tools to pick up," said Richard Mann, PhD, a principal investigator at Columbia's Mortimer B. Zuckerman Mind Brain Behavior Institute and a senior author of the paper. "But our new algorithm lets us sweep through these millions of lines of genetic code and pick up even the faintest signals, resulting in a much more complete picture what DNA encodes."

Geneticists have long looked for ways to decipher the mysteries hidden in DNA. One such mystery has involved a particularly pervasive class of genes known as the Hox genes.

"Hox genes are the body's master architects; they drive some of the earliest and most critical aspects of growth and differentiation, such as where in a developing embryo the head and limbs should be positioned," said Dr. Mann, who is also the Higgins Professor of Biochemistry and Molecular Biophysics (in Systems Biology) at Columbia University Irving Medical Center. "Hox genes do this by producing proteins called transcription factors, which bind to DNA sequences in order to turn large cohorts of genes on or off; like flipping thousands of switches in exactly the right order."

But decades of research into Hox genes uncovered a paradox: Even though each individual Hox gene guided a *different* feature of growth, the Hox transcription factors were all binding strongly and visibly to the *same* set of easily identifiable DNA sequences.

In 2015, Dr. Mann and his team discovered that the Hox transcription factors were *also* binding at many other locations — just more discretely at so-called ‘low-affinity sites.’ The scientists believed these low-affinity binding sites to be key to the Hox transcription factors being able to drive one aspect of development versus another. The problem remained how to decipher these sites from the genome.

To address this challenge, Dr. Mann and his lab joined forces with the lab of Harmen Bussemaker, PhD, a Professor in Columbia’s Department of Biological Sciences and Systems Biology and an expert in building computational models of genetic activity.

A few years ago, the two labs developed a genetic sequencing method called SELEX-seq to systematically characterize all Hox binding sites. But their approach still had limitations: It required the same DNA fragment to be sequenced over and over again. With each new round, more pieces of the puzzle were revealed, but information about those critical low-affinity binding sites remained hidden.

“It was akin to running the same paragraph through Google translate multiple times, but in the end still only ten percent of the words are accurately translated,” said Dr. Mann.

To overcome this challenge, Dr. Bussemaker and his team developed a sophisticated new computer algorithm that was able to explain — for the first time — the behavior of all DNA sequences in the SELEX-seq experiment. They called this algorithm *No Read Left Behind*, or *NRLB*.

“In simple terms, *NRLB* allows us cover the entire spectrum of binding sites — from the highest to the lowest affinity — with a much greater degree of sensitivity and accuracy than any existing method, including state-of-the-art deep learning algorithms” said Dr. Bussemaker, who was the paper’s other senior author. “Building on that foundation, we now hope to develop more in-depth biological and computational models to help answer the most complicated questions about the genome.”

“For example, diseases such as schizophrenia, Parkinson’s disease and autism have been mapped to particular DNA regions that do not appear to have a clear function,” said Dr. Mann. “With *NRLB*, scientists could potentially piece together how transcription factors bind to and activate those regions. This will be critical for finding ways to manipulate that activity to one day reduce one's risk of disease."

###

This paper is titled: "Accurate and sensitive quantification of protein-DNA binding affinity." Additional contributors include first author Chaitanya Rastogi, H. Tomas Rube, PhD, Judith Kribelbauer, Justin Crocker, PhD, Ryan Loker, Gabriella Martini, Oleg Laptenko, PhD, William Freed-Pastor, MD, Carol Prives, PhD, and David Stern, PhD.

This research was supported by the National Institutes of Health (R01HG003008, R35GM118336, CA87497, CA196234, T32GM08279707, G20RR030893) and the Howard Hughes Medical Institute.

The authors report no financial or other conflicts of interest.

Columbia University's ***Mortimer B. Zuckerman Mind Brain Behavior Institute*** *brings together a group of world-class scientists and scholars to pursue the most urgent and exciting challenge of our time: understanding the brain and mind. A deeper understanding of the brain promises to transform human health and society. From effective treatments for disorders like Alzheimer's, Parkinson's, depression and autism to advances in fields as fundamental as computer science, economics, law, the arts and social policy, the potential for humanity is staggering. To learn more, visit: zuckermaninstitute.columbia.edu.*