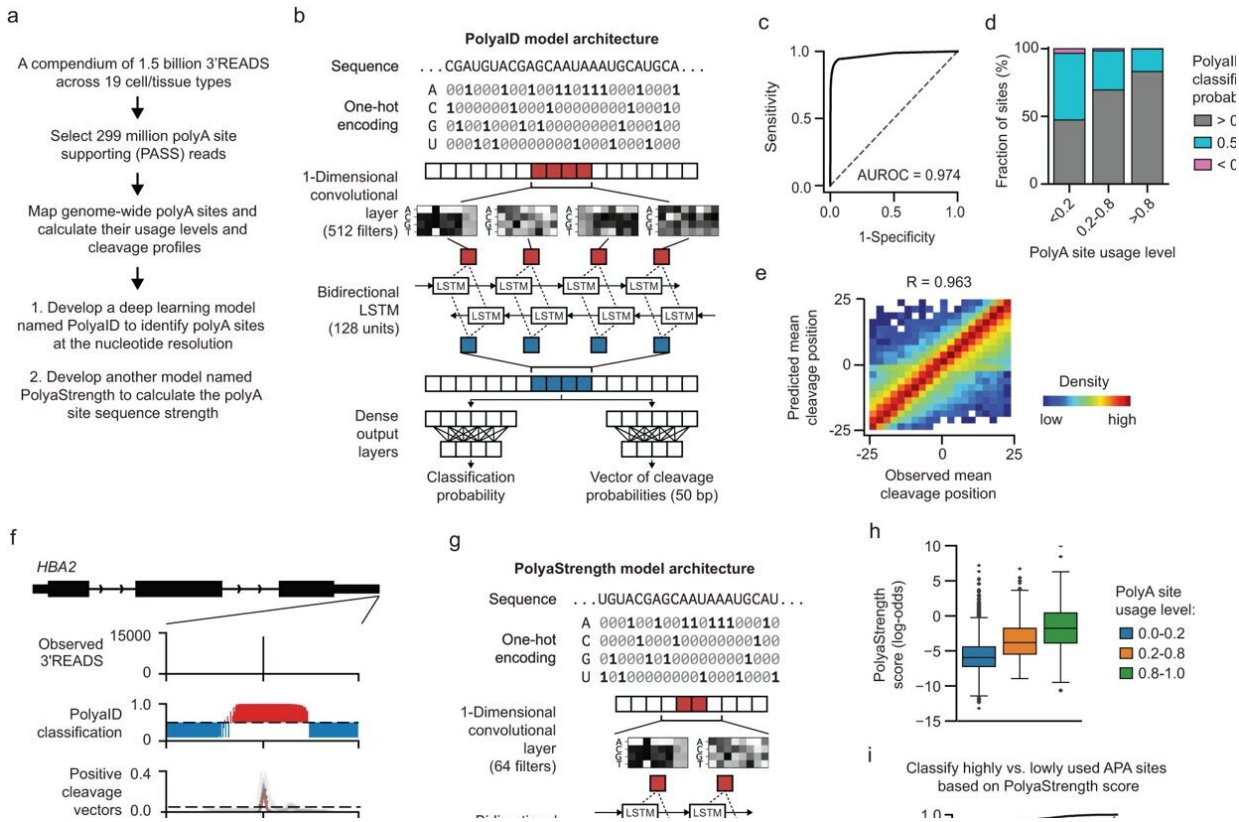
Developing deep learning models to understand the human genome

December 15 2023, by Olivia Dimmer



Developing deep learning models to identify polyA sites at nucleotide-level resolution and calculate polyA site strength. Credit: *Nature Communications* (2023). DOI: 10.1038/s41467-023-43266-3

Northwestern Medicine scientists have developed a deep learning algorithm capable of identifying the location where a genetic process

called polyadenylation occurs on the genome, according to findings published in *Nature Communications*. Investigators say the development has the potential to accelerate research around diseases and disorders that occur when the process of DNA transcription goes awry.

Polyadenylation, a crucial step in normal gene expression, is the process through which nucleotides are added to RNA, stabilizing and readying them for translation into proteins. Polyadenylation is also essential for telling RNA when to halt transcription, an event known as termination, which prevents erroneous gene expression.

Although the process has been implicated in cancer, neonatal diabetes and some genetic disorders, little is known about the location of polyadenylation sites on the genome and the factors that might influence it.

"The process we are focusing on in my lab is understanding how this termination is controlled," said Zhe Ji, Ph.D., assistant professor of Pharmacology and at the McCormick School of Engineering, and senior author of the study.

"This polyadenylation process is essential for defining the three-prime ends of the RNA and also for transcription termination. In my field, there's this challenging question: 'Where are these polyadenylation sites localized in the genome sequence?' We want to understand where and how these signals occur so we can optimize them across the human genome."

In the current study, Emily Stroup, a Ph.D. candidate in the Driskill Graduate Program who was first author of the study, developed deep-learning models to analyze and predict where polyadenylation (or polyA) sites occur on the human genome, how DNA cleavage occurs around those sites, and the strength of those polyA sites in relation to others in

the same gene.

"The major advantage of having a series of models like this is that we can learn from what the model learned during training to better understand of how polyA site usage is regulated and the factors that determine where the exact cleavage sites are chosen," Stroup said.

Using the model, investigators found that polyadenylation sites are influenced by a variety of signals expressed near the site. The model also allowed investigators to identify human genome sequences in which polyadenylation occurs properly and most efficiently, providing a road map for future research.

"We were able to look at the factors that control how precisely cleavage sites are defined and then what controls the usage of those sites and how we can modulate that," Stroup said.

"We were also able to see where the polyA sites are located in the genome, how they're positioned within the gene relative to both other polyA sites and the overall landscape of the surrounding genome, to better understand these external genomic factors that control if a site's used or not. This is something that we were able to do on a large scale that hasn't been done before."

By understanding how and why polyadenylation occurs where it does on the human genome, investigators may be able to develop therapeutic approaches for correcting the process in the context of disease, Ji said.

"It's a breakthrough, basically. With this model, for the first time, we can achieve this single-nucleotide resolution prediction of these polyA sites across the human genome because the site is determined by multiple signals," Ji said.

Moving forward, Stroup and Ji are focusing on developing similar models for other species, including zebrafish, fruit flies, and yeast, to compare the location of polyA sites in the genomes of different animals.

"In early results, we've found a lot of signal differences across different species, and that will help us understand the evolution of these signals across species," said Ji, who is also a member of the Robert H. Lurie Comprehensive Cancer Center of Northwestern University.

"We're hoping that will allow us to understand how signal mutations can contribute to population genetics or different kinds of human diseases, such as muscular dystrophy, neuronal disorders and cancers."

More information: Emily Kunce Stroup et al, Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease, *Nature Communications* (2023). DOI: 10.1038/s41467-023-43266-3

Provided by Northwestern University